Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **00**, 1–23
https://doi.org/10.1093/jrsssc/qlad003

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

On the competitive facility location problem with a Bayesian spatial interaction model

Shanaka Perera[1], Virginia Aglietti[1,2] and Theodoros Damoulas[1,2]

[1]Department of Computer Science, University of Warwick, Coventry, UK
[2]The Alan Turing Institute, London, UK

Address for correspondence: Shanaka Perera, Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK. Email: msb15sp@mail.wbs.ac.uk

Abstract

The competitive facility location problem arises when businesses plan to enter a new market or expand their presence. We introduce a Bayesian spatial interaction model which provides probabilistic estimates on location-specific revenues and then formulate a mathematical framework to simultaneously identify the location and design of new facilities that maximise revenue. To solve the allocation optimisation problem, we develop a hierarchical search algorithm and associated sampling techniques that explore geographic regions of varying spatial resolution. We demonstrate the approach by producing optimal facility locations and corresponding designs for two large-scale applications in the supermarket and pub sectors of Greater London.

Keywords: Bayesian spatial interaction model, competitive facility location, multiresolution, optimisation problem, spatial data

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. In Great Britain, online sales as a proportion of total retail sales have tripled in a decade, reaching 21% in 2019 (ONS, 2020). Therefore, it is essential to understand how customers interact with physical business facilities in order to design new commercial centres in competitive markets. We propose a modelling framework that accounts for customer behaviour to identify the optimal criteria for a company to enter a new market or expand its presence in a geographical region. We aim to address three of the most pivotal question 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 (Harold, 1929). 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 (Murray, 2018; Toregas et al., 1971). In contrast, *competitive facility location problems* emphasise industries such as retail businesses and commercial services, which consider competition among stores when choosing their sites (Berman et al., 2009; Drezner, 2014). These companies compete to attract customers buying power in a given area to capture market share.

One of the earliest probabilistic approaches for estimating market share was proposed by Huff (1963) based on the gravity model (Reilly, 1931). Huff's formulation states that the value or utility

Received: December 6, 2021. **Revised:** July 18, 2022. **Accepted:** December 20, 2022

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 explain customer behaviour (Drezner, 2006; Wilson, 1971). Customers are assumed to patronise shopping centres based on their satisfaction indicated by a utility function (Drezner, 2014). The competitive location facility problem integrates the spatial interaction between customers and stores into the optimisation model according to utility models (Benati & Hansen, 2002; Freire et al., 2016).

Inspired by the literature on gravity models, we develop 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. We model the probability of a customer visiting each facility in a region through Gaussian densities in geographic space. Specifically, each density is centred on a facility with variance that is further determined by its attractiveness which in turn 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 we assume to be a function of their socio-demographic characteristics. 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, we integrate such demand elasticities by adding dummy facilities as proposed by Leonardi and Tadei (1984) and Drezner and Drezner (2012). We adopt a Bayesian approach (van de Schoot et al., 2021) that enables us to adequately account for the uncertainty associated with the customer interactions with the facilities. Our framework not only gives accurate predictions but also 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. We address this issue by resorting to variational inference (Jordan et al., 1999)

We adopt the BSIM method to model customer behaviour and estimate revenue generated at the new stores. Our 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 (Drezner et al., 2012; Küçükaydin et al., 2011). We 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 given set of potential sites and structures, in terms of store characteristics, within a set budget.

In the process of establishing new facilities, the users are unable to provide an exhaustive set of potential sites, or this set is too large that it becomes computationally expensive. We propose a hierarchical search method that starts with a broad area and narrow the search to several regions to explore the neighbouring locations using a quadtree approach. The initial set of candidates are formed ensuring that more potential sites are situated in areas with a high-density ratio between customer purchasing power and existing facilities. We adopt a non-parametric approach, kernel density estimation, to estimate the probability density functions (Bishop, 2006). According to the density ratio, the samples are generated from a multiresolution grid structure (Samet, 2006) and an inhomogeneous Poisson point process (Lewis & Shedler, 1979). We evaluate the performance of these methods and regular grid sampling using synthetic experiments and demonstrate that the multiresolution grid structure outperforms other approaches.

In the literature, the applications for competitive facility location problems are limited to synthetic experiments and, real-world applications are restricted to small regions or applied at the aggregate level because access to large scale spatial data is usually expensive (Benati & Hansen, 2002; Drezner & Drezner, 2012; Freire et al., 2016). In contrast, we develop a granular level spatial dataset with supermarket characteristics for Greater London by utilising data from commercial and open sources. We demonstrate the BSIM and estimate the parameters to derive customer interactions with supermarkets in London. Two real-world applications are presented to identify optimal facility locations: establishing new supermarket stores for a couple of chains and a new company entering the pub market in London.

Our main contributions are (a) we develop a Bayesian spatial interaction model (BSIM) that can be used to make probabilistic predictions of revenues or demand generated at business facilities and formulate the relationship between distance and attractiveness of facilities jointly, using a facility-specific probability distribution; (b) we 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) we propose a search algorithm based on the quadtree method to explore geographic regions of varying spatial resolution hierarchically; and (d) we demonstrate the optimal facility locations and their designs to establish new stores in Greater London for two industries.

To the best of our knowledge, we are the first to present a fully integrated competitive facility problem that includes both the spatial interaction modelling component and the store location optimisation framework that was demonstrated in one of the major cities in the world using a large-scale dataset with over 1,000 supermarkets, 1,500 pubs, and 150,000 customer regions.

The paper is organised as follows. In the next section, we introduce the BSIM and formulate the optimisation problem. In Section 3, we demonstrate and evaluate the optimal location search using synthetic experiments. In Section 4, we introduce a comprehensive spatial dataset. Next, in Section 5, we demonstrate the optimisation framework using a couple of real-world applications. Finally, conclusions and future research directions are discussed in Section 6.

2 Methodology

In this section, first we introduce the Bayesian spatial interaction model (BSIM). Next, we introduce the competitive facility location problem and a framework to search for optimal sites.

2.1 Bayesian spatial interaction model (BSIM)

Suppose there are N customers and the nth customer is residing in location $\mathbf{m}_n \in \mathbb{R}^2$ having sociodemographic 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 $\phi_s \in \mathbb{R}^D$ in a bounded region τ. 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 $\psi(\boldsymbol{\mu}_s,\ \boldsymbol{\Sigma}_s)$, centred 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 nth customer to visit the sth store,

$$u_{ns} = \psi(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \begin{cases} \dfrac{\exp\left(-d_{ns}^2/2\sigma_s^2\right)}{2\pi\sigma_s^2\left(1-\exp\left(-d_T^2/2\sigma_s^2\right)\right)}, & 0 \le d_{ns} \le d_T, \\ 0, & \textit{otherwise}, \end{cases} \tag{1}$$

where d_{ns} denotes 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. For illustration purposes, consider three stores where each has a truncated Gaussian distribution centred on the store, as shown in Figure 1.

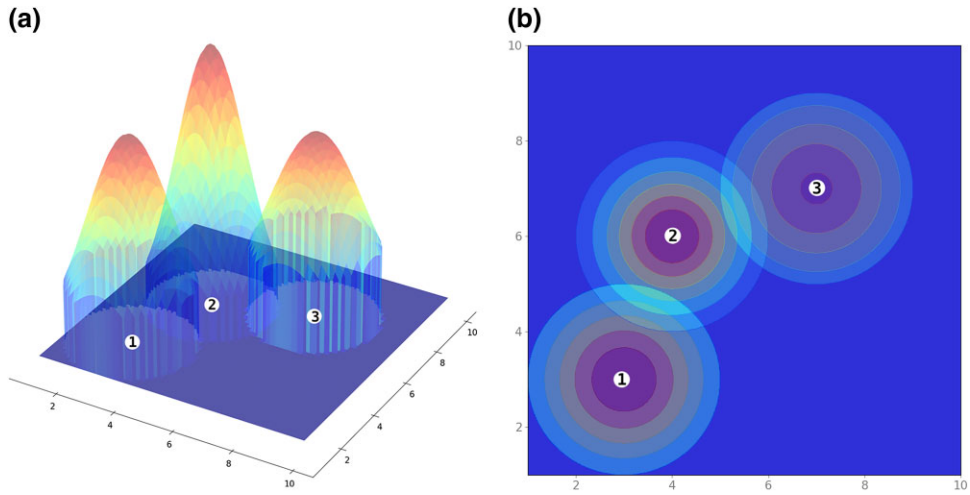

Figure 1. Illustration of the truncated Gaussian centred on three sample stores: (a) 3D visualisation and (b) 2D visualisation. The white dots indicate the store location. There is a hard border around the distributions beyond which the PDF is equal to zero.

Variance of the Gaussian σ_s^2 is formulated as a function of store characteristics ϕ_s and the non-observable store characteristics $\varepsilon_s \in \mathbb{R}$,

$$\sigma_s^2 = \exp(\lambda^\top \phi_s + \varepsilon_s), \tag{2}$$

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

$$p_{ns} = \frac{u_{ns}}{\sum_{j=1}^{S} u_{nj}}. \tag{3}$$

Note that we 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 which falls in the interval of [0, 1]. 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. We further introduce lost demand as proposed by Leonardi and Tadei (1984) and Drezner and Drezner (2012). The lost demand is assumed to be attracted by a dummy facility which is treated as an additional competing facility. Henceforth, we advance the model by introducing utility term u_{nd} assuming a dummy facility in addition to the existing alternatives,

$$p_{ns} = \frac{u_{ns}}{\sum_{j=1}^{S} u_{nj} + u_{nd}}. \tag{4}$$

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

$$p_n = \sum_{s=1}^{S} p_{ns} = \frac{\sum_{n=1}^{S} u_{ns}}{\sum_{j=1}^{S} u_{nj} + u_{nd}} \leq 1. \tag{5}$$

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. We set the variance of the Gaussian placed on the dummy facility as $\sigma_d^2 = d_T/4$ 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

$$u_{nd} = \begin{cases} \dfrac{\exp\left(-d_D^2/2\sigma_d^2\right)}{2\pi\sigma_d^2\left(1-\exp\left(-d_T^2/2\sigma_d^2\right)\right)}, & 0 \le d_D \le d_T \\ 0, & \textit{otherwise.} \end{cases} \tag{6}$$

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

$$g_n = \beta^\top \mathbf{v}_n. \tag{7}$$

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

$$r_s = \sum_{n=1}^{N} g_n p_{ns}. \tag{8}$$

Finally, the complete data likelihood is

$$p(\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \varepsilon, \sigma^2) = \prod_{s=1}^{S} \mathcal{N}\left(y_s \middle| \sum_{n=1}^{N} \boldsymbol{\beta}^\top \mathbf{v}_n \frac{\psi(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)}{\sum_{j=1}^{S} \psi(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + u_{nd}}, \sigma^2\right), \tag{9}$$

with $\mathbf{Y} = \{y_1, \ldots, y_S\}$ and the model assumes constant-variance (σ^2) for the Gaussian noise. The graphical representation for the BSIM is presented in Figure 2.

2.1.1 Prior distributions

We assign prior distributions to all model parameters. First, we define a hierarchical prior distribution for β, which we assume to be a Gaussian with mean $\boldsymbol{\mu}_\beta$ and covariance $\alpha^{-1}\mathbf{I}$,

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

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

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

Similarly, we assign a Gamma prior distribution with shape ρ_1 and scale ρ_2 for the likelihood precision parameter γ,

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

Finally, the following Gaussian prior distributions are selected for $\boldsymbol{\lambda}$ and $\boldsymbol{\varepsilon}$,

$$p(\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\lambda}; \boldsymbol{\mu}_\lambda, \boldsymbol{\varrho}_\lambda \mathbf{I}),$$
$$p(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon}; \boldsymbol{\mu}_\varepsilon, \boldsymbol{\varrho}_\varepsilon \mathbf{I}).$$

2.1.2 Posterior distribution

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

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}, \tag{10}$$

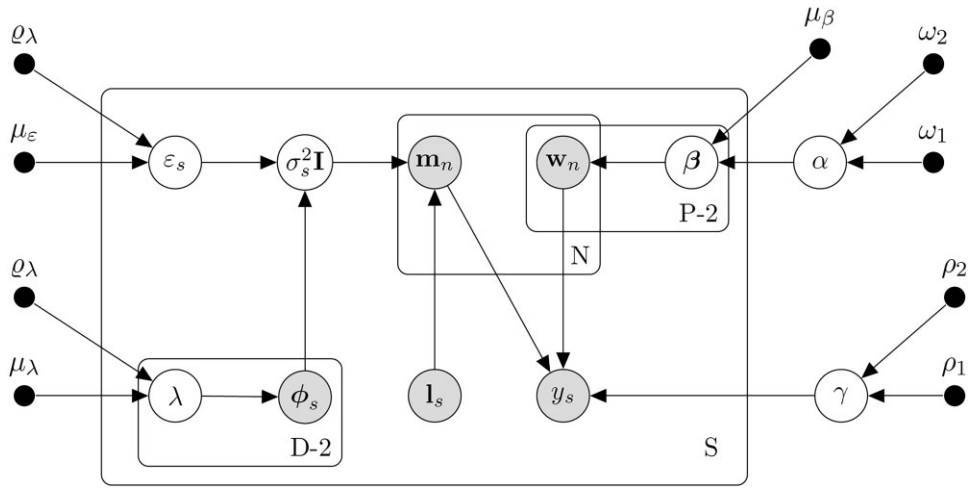

Figure 2. Plate diagram for the graphical representation for the BSIM. Specifically, this express the spatial interaction between S number of stores with each store revenue y_s, located at $\mathbf{l}_s$ with store features $\boldsymbol{\phi}_s$ and N number of customers located at m_n with P-2 characteristics $\vec{w}_n$. We use Gaussian distributions as priors for $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, ε 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.

where the marginal density takes the form

$$p(\mathcal{D}) = \int \cdots \int p(\mathcal{D}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \gamma) p(\boldsymbol{\beta}|\alpha) p(\alpha) p(\boldsymbol{\lambda}) p(\boldsymbol{\varepsilon}) p(\gamma)\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\alpha\, \mathrm{d}\boldsymbol{\lambda}\, \mathrm{d}\boldsymbol{\varepsilon}\, \mathrm{d}\gamma. \tag{11}$$

2.1.3 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 [equation (11)], we resort to approximate inference by employing scalable Variational Inference (VI) (Jordan et al., 1999). The details of this are presented in the Supplementary Material. We assume customers make their choices according to the BSIM, and the estimated posterior parameters are used for the optimisation problem in locating new facilities.

2.2 Optimal facility location

We 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. We 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 in a bounded region τ. 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 is 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 ϕ_{lr}. Thus, we obtain the truncated Gaussian PDF by

$$\mathbf{Z}_{lrn} = \begin{cases} \dfrac{\exp\left(-d_{nl}^2 / 2\exp\left(\lambda^\top \phi_{lr}\right)\right)}{2\pi \exp\left(\lambda^\top \phi_{lr}\right)\left(1 - \exp\left(-d_T^2 / 2\exp\left(\lambda^\top \phi_{lr}\right)\right)\right)}, & 0 \le d_{nl} \le d_T \\ 0, & \textit{otherwise}. \end{cases} \tag{12}$$

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

$$u_{nl} = \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}. \tag{13}$$

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

$$p_{nl} = \frac{u_{nl}}{u_{nL} + u_{nd} + \sum_{l' \in \tilde{L}} u_{nl'}}, \tag{14}$$

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

$$y_{L^*} = \sum_{l \in \tilde{L}} \sum_{n=1}^{N} g_n \frac{u_{nl}}{u_{nL} + u_{nd} + \sum_{l' \in \tilde{L}} u_{nl'}}. \tag{15}$$

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

$$\sum_{l \in \tilde{L}} \sum_{r=1}^{R} c_{lr} x_{lr} \leq B. \tag{16}$$

2.2.1 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, we formulate three unique objective functions denoted by $v(x_{lr})$.

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

$$\sum_{n=1}^{N} g_n \frac{\sum_{l \in \tilde{L}} \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}}{u_{nL} + u_{nd} + \sum_{l \in \tilde{L}} \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}}. \tag{17}$$

Case II: Suppose a company already has a chain of existing facilities in a market $\hat{L} \subset L$ 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

$$\sum_{n=1}^{N} g_n \frac{\sum_{l' \in \hat{L}} u_{nl'} + \sum_{l \in \tilde{L}} \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}}{u_{nL} + u_{nd} + \sum_{l \in \tilde{L}} \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}}. \tag{18}$$

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

$$\sum_{n=1}^{N} g_n \frac{u_{nL} + \sum_{l \in \tilde{L}} \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}}{u_{nL} + u_{nd} + \sum_{l \in \tilde{L}} \sum_{r=1}^{R} \mathbf{Z}_{lrn} x_{lr}}. \tag{19}$$

2.2.2 *Optimisation problem*

Given the above definitions, we 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,

$$\max_{x_{lr}} \quad v(x_{lr}) \tag{20}$$

$$\text{subject to:} \quad \sum_{l\in\tilde{L}}\sum_{r=1}^{R} c_{lr}x_{lr} \le B \tag{21a}$$

$$\sum_{l\in\tilde{L}}\sum_{r=1}^{R} x_{lr} \le k \tag{21b}$$

$$\sum_{r=1}^{R} x_{lr} \le 1 \quad \text{for} \quad l \in \tilde{L} \tag{21c}$$

$$x_{lr} \in \{0, 1\}, \quad \text{for} \quad r = 1\ldots R; \quad l \in \tilde{L}, \tag{21d}$$

where constraint (21a) is an upper limit for the total cost, (21b) limits the maximum number of facilities, and (21c) 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 (Kellerer et al., 2004), 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 we use 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 (Wolsey & Nemhauser, 1999). We solve the optimisation problem using constraint programming with the CP optimiser on IBM ILOG CPLEX studio 20.1.

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. We propose a hierarchical search algorithm 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 1.

We present three options to generate the initial set of candidate locations for the hierarchical search algorithm, as discussed in the following sections. We split the potential facility locations into random samples in the first level before executing the optimisation algorithm. Decomposing the larger matrix into smaller samples improves computational complexity in optimisation algorithms. Additionally, 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, we 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.

Algorithm 1: 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**
  $L^*, v \leftarrow$ findOptimalLocs(*samples*, B, k);
  save L^* in $\mathcal{L}$;
end
$\Delta v \leftarrow \tau$;
$v \leftarrow 1$;
while $\Delta v \geq \tau$ **do**
  $v_0 \leftarrow v$;
  $L^*, v \leftarrow$ findOptimalLocs($\mathcal{L}$, B, k);
  $\mathcal{L} \leftarrow$ getQuadtree(L^*);
  $\Delta v \leftarrow (v - v_0)/v_0$;
end

We propose three sampling methods 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.

2.3.1 *Regular grid sampling*

The regular grid sampling method does not account for the customer and facilities' spatial variability; thus, the candidate locations are distributed at regular distance 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. We create a set of random samples 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.

2.3.2 *Density-based sampling*

We propose a density-based sampling method to create the initial candidate locations to overcome sampling errors in regular grid sampling. We adopt a non-parametric approach, kernel density estimation, for estimating the probability density function (Bishop, 2006). 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

$$f_s(x) = \frac{1}{Sh} \sum_{s=1}^{S} K\left(\frac{x - \mathbf{l}_s}{h}\right), \tag{22}$$

where $K(\cdot)$ is a kernel function, we choose Gaussian kernel with band-width parameter h, optimally selected according to Silverman (1986). The spending power g_n (equation 7) is unevenly distributed at customer locations $\mathbf{m}_n$. Hence, we consider a weighted kernel density estimator to model customer spending capacity (Gisbert, 2003). The spending capacity g_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:

$$f_n(x) = \frac{1}{Nh} \sum_{n=1}^{N} w_n K\left(\frac{x - \mathbf{m}_n}{h}\right) \tag{23}$$

We calculate a ratio $f_r(x)$ between the density estimates, which provides an indicator of how dense the area is in terms of customers spending power compared to the available facilities,

$$f_r(x) = \frac{f_n(x)}{f_s(x)}. \tag{24}$$

We propose two sampling methods using the estimated density ratio.

Sampling with inhomogeneous Poisson point process

We simulate the potential set of locations 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 (Cressie, 1994). In contrast, the intensity function of an inhomogeneous Poisson process is a nonconstant function $\lambda(x)$ of spatial location $x \in \mathcal{R}^2$.

We simulate IPPP through Lewis and Shedler (1979) thinning algorithm. First, we obtain a random number η^* from a Poisson distribution with mean $\mu(A) = \int_A \lambda(x)\,dx$. Next, we simulate a homogeneous Poisson point process with intensity value λ^* which is an upper bound of the intensity function $\lambda(x)$. For this, we use the maximum of the ratio between the density estimates, $\lambda^* = \max f_r(x)$. 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 locations for the hierarchical search. The second level of the hierarchical search does not continue recursively since we are not using the grids 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, we estimate $f_r(x)$ 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 2.

3 Simulation study

We design a simulation study to experiment with the optimisation problem using the three objective functions introduced and compare the performance using the three sampling methods proposed in the Section 2. We also compare the computational performance of the methods by observing the run time of each optimisation problem. All the experiments are executed on a Intel Core i5 CPU (2.3 GHz Dual-Core and 8 GB of RAM). Furthermore, a simulation study is presented in the Supplementary Material to examine the inferences obtained for BSIM under different synthetic settings characterised by an increasing number of stores and customers.

First, we simulate the data from a spatial process that closely matches the extended BSIM framework introduced in Section 2 with the dummy facilities. The process is defined as

$$y_s \,|\, \beta, \boldsymbol{\lambda}, \sigma^2 \sim \mathcal{N}(r_s, \sigma^2), \tag{25}$$

where we assume the reasonable distance a customer is willing to travel is 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 Figure 3a. The customers' satisfied demand (p_n) from the existing stores are shown in Figure 3b. The store locations are randomly sampled, and their current revenue is displayed in Figure 3c. We assume two possible designs ($r = 2$), say small and large facility structures, are 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.

Algorithm 2: Multiresoltion grid structure

$x \leftarrow$ constructPointsMeshgrid(*region*);
grid $\leftarrow$ decompose(*region*, m) ; // decompose *region* into m sub-blocks
foreach c_i *in grid* **do** // for each cell c_i in the *grid*
  $\mu_r \leftarrow \text{mean}(f_r(x))$; // Mean of $f_r(x)$ of points in cell c_i
  save μ_r in c_i;
end
$\hat{q} \leftarrow \max(\mu_r)/q$; // q denotes the depth of resolution
foreach c_i *in grid* **do**
  points $\leftarrow$ midpoint(c_i) ; // Midpoints of cell c_i
  save *points* in *out*;
  for $j = 1$ **to** q **do**
    if $(j-1) \times \hat{q} \geq \mu_r \leq j \times \hat{q}$ **then**
      for $g = 1$ **to** j **do**
        $c_i \leftarrow$decompose(c_i, 4) ; // Decompose c_i into 4 square submatrices and repeat recursively
        points $\leftarrow$ midpoint(c_i);
        save *points* in *out*;
      end
      exit
    end
  end
end

3.1 Demonstration of the optimal facility location with varying objective functions

Given the above synthetic setting, we solve the optimisation problem to find the optimal location for one new facility with a budget of ten($B = 10$) for the three objective functions discussed in Section 2. We use a regular grid sampling approach to generate the potential facility locations, as presented in Figure 3c. The results of the optimisation problem with the objective function in case I (equation 17), 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 (Figure 4a). In case II (equation 18), a franchise opening a new facility, the optimal location moves away from the other facilities in the chain as displayed in Figure 4b. Revenue of the new facility reduces compared to the case I, but the total sales of the chain facilities are increased, as shown in Table 1. Finally, in case III (equation 19), when opening a new facility in a monopolistic market, the new store locates away from all the existing facilities (Figure 4c) 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 1). In all three cases, the optimal facility design is large size.

3.2 Evaluation of sampling methods for the hierarchical search

We experiment with the hierarchical search using the three sampling methods proposed in Section 2. The synthetic setting remains as described above. Experiments are performed with the objective function where a new company is entering the market (equation 17) 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.

3.2.1 Regular grid sampling

We use a regular grid with dimensions of 15×15 to generate the initial candidate locations. The locations are split into four samples using the stratified sampling method. The optimisation

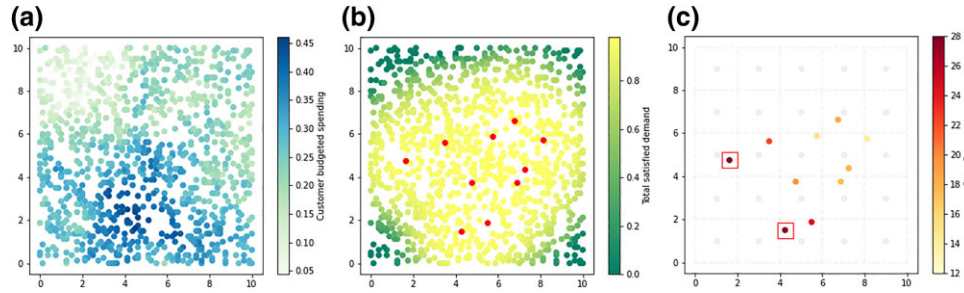

Figure 3. (a) Simulated customer locations ($N = 1,000$) and budgeted spending (colour gradient). (b) Satisfied customer demand (colour gradient) and the existing stores (red). (c) Revenue of the existing stores (colour gradient) and potential store locations (grey).

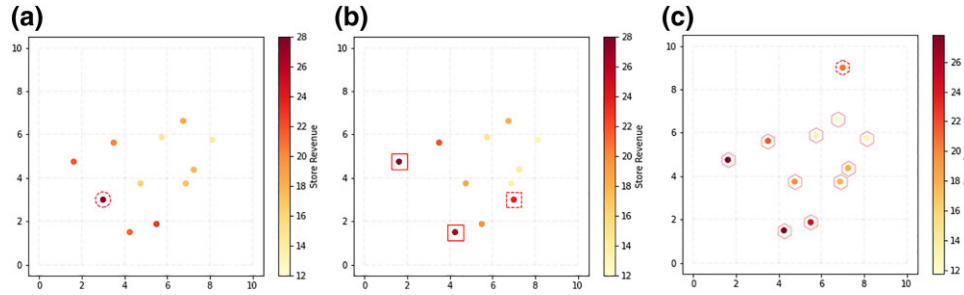

Figure 4. The experiment is to find the optimal location for a new facility under three different objectives. (a) Maximise the revenue of the new facility (equation 17). (b) Maximise the revenue of all facilities owned by the franchise (equation 18). Square indicates the existing facilities owned by the franchise. (c) Maximise the revenue of all facilities in the market (equation 19). Hexagons indicate that all facilities owned by the same company. The optimal location is shown within the red colour dashed circle, square, and hexagon.

Table 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

problem is solved independently for each sample to identify two optimal sites forming eight in total, as shown in Figure 5a. For the next level of the hierarchical search, the neighbourhood locations are produced using the quadtree method forming 40 potential sites as reported in Figure 5b. The recursive algorithm stops after two iterations producing two locations to establish the new facilities with the two designs.

3.2.2 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. Figure 6 demonstrates the density contour plot generated for customer purchasing power, store locations, and the ratio of the two density estimates in the area. We use a 100×100 meshgrid to estimate the density and calculate the ratio.

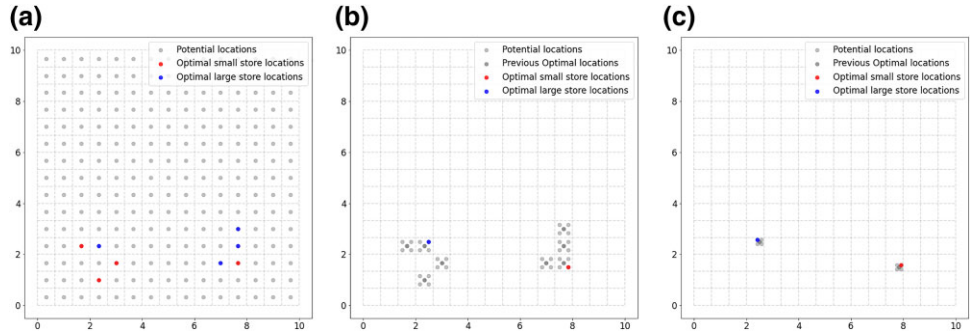

Figure 5. Visual progression for regular grid sampling for hierarchical search. (a) Initial candidate locations generated from 15 × 15 regular grid. Eight optimal locations are found from each sample producing one small and large design facilities. (b) Neighbourhood locations for the optimal locations generated from the previous step and the new optimal locations. (c) Final optimal locations are derived from 10 potential locations.

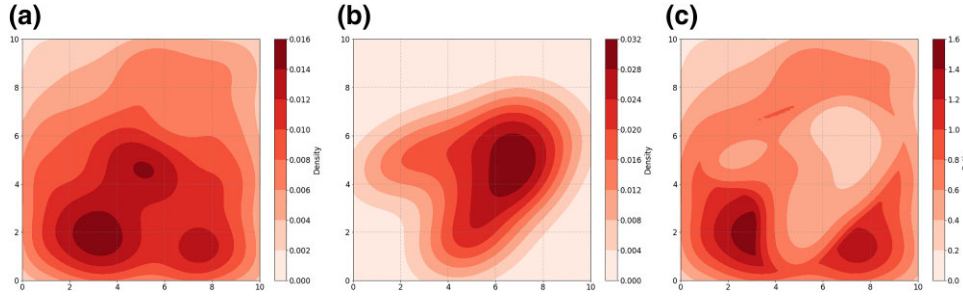

Figure 6. Demonstrates the density contour plot generated for (a) customer spending, (b) store locations, and (c) ratio of the two density estimates in the area.

Sampling with inhomogeneous Poisson point process: We use the maximum estimated ratio from the meshgrid 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 Figure 7. These eight locations become the potential sites for the final iteration to find the optimal facilities.

Multiresolution sampling: We create a regular grid of 5×5 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 three-quantiles of the average values to decide the resolution of each cell. Figure 8a presents the multiresolution samples used as the potential locations. The set of candidates are split into four random samples and solve the optimisation problem independently. The algorithm stops after two iterations providing the optimal facilities, as shown in Figure 8c.

3.2.3 Comparison between sampling methods

We compare the results 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 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.

We extend the comparison by simulating the experiment 1,000 times. We create 1,000 datasets 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

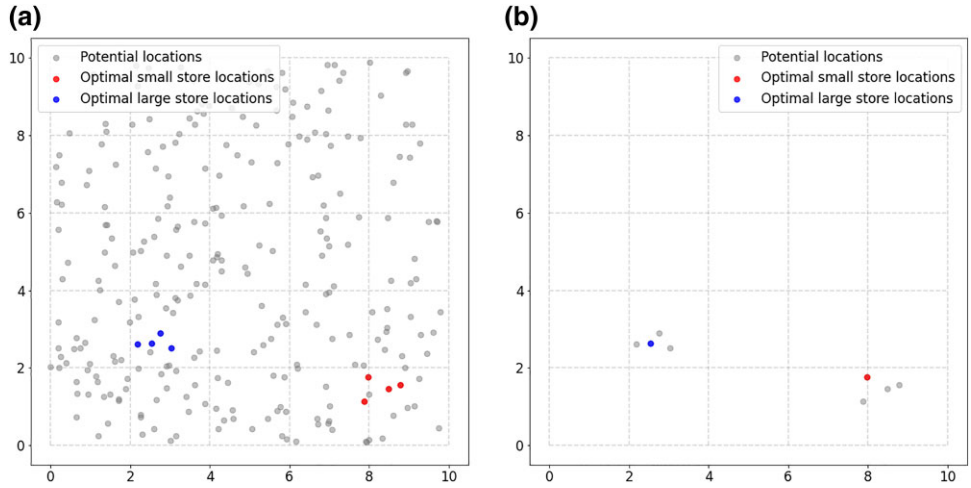

Figure 7. Visual progression for IPPP sampling for hierarchical search.(a) Random samples are generated from the IPPP as the potential locations for the optimisation problem. Eight optimal facilities are found, with each sample producing one small and large facility location. (b) The optimal locations of the previous stage becomes the candidate sites for the second level from which the optimal locations are identified.

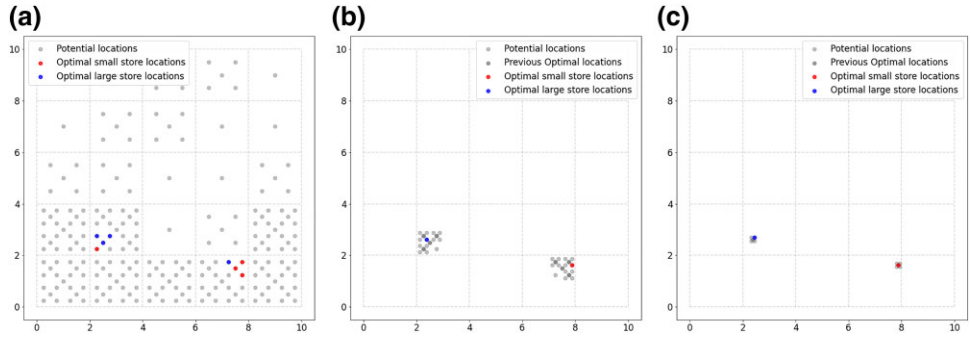

Figure 8. The progression of the multiresolution sampling method to find the optimal locations. (a) Multiresolution samples for 5×5 grids with a depth of three. (b) Neighbourhood locations for the optimal sites generated from the previous step and the new optimal locations. (c) Final optimal locations are derived from 10 potential sites.

Table 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 = 15×15)	53.83	231
IPPP	271 (samples = 4)	53.78	240
Multiresolution	217 (depth = 3)	**53.85**	275

experimented at two levels to demonstrate the behaviour based on the initial sample size. We compare the performance in Table 3.

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 values 55% and 60% of the time with the varying sample sizes. With the increase in the starting number of candidate locations, the comparison

Table 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 = 8 × 8)	434	59.76	33
	225 (grids = 15 × 15)	390	59.85	252
IPPP	76 (samples = 1)	13	58.84	57
	262 (samples = 4)	12	59.62	288
Multiresolution	73 (depth = 2)	553	59.81	38
	217 (depth = 3)	**598**	**58.89**	241

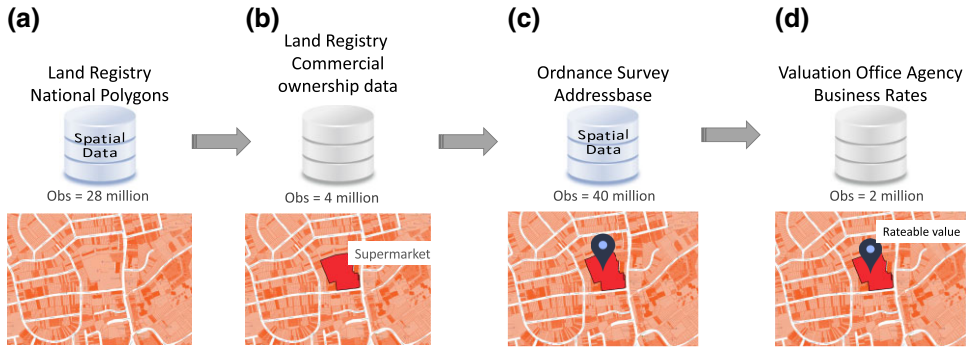

Figure 9. The flow diagram presents the steps in developing a dataset of supermarkets in the UK with their geo-location coordinates. (a) The map shows title polygons. (b) Filter only the titles owned by supermarket chains. (c) Spatially join to identify the data from OS. (b) Finally, join with VOA data to get only the supermarkets and their rateable values.

between the regular grid and multiresolution methods is broadened. IPPP has performed comparatively poorly with only less than 2% of the experiments 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, we have not considered developing a method to explore the neighbouring sites. We can conclude that the multiresolution sampling method could produce better results with a low number of starting locations while being efficient.

4 Geospatial dataset of supermarkets in Greater London

We develop a large scale geospatial dataset for supermarket chains in the UK using multiple data sources. To the best of our knowledge, this is the first study utilising these datasets together to create a granular level supermarket dataset with its characteristics. Additionally, we use the customer level and pubs datasets introduced by Perera et al. (2021).

First, we filter the properties owned by the leading supermarket chains (Asda, Co-op Food, Iceland, Lidl, Marks & Spencer, Morrisons, Sainsbury's, Tesco, Waitrose) from the commercial and corporate ownership data by HM Land Registry (2020b). The ownership dataset provides details on registered titles in England and Wales owned by UK companies. However, the filtered data contain other types of businesses owned by the respective supermarket chains, such as their warehouses. Valuation Office Agency (2019) data provide the categorisation of the non-domestic properties along with their rateable values and floor sizes. We filter the VOA dataset to extract the properties representing supermarket or food store categories. Since there is no direct link between the two datasets, we join the VOA data with the Ordnance Survey (2019) Addressbase data using the cross-reference to obtain the geo-coordinates of the properties. Next, we join the filtered

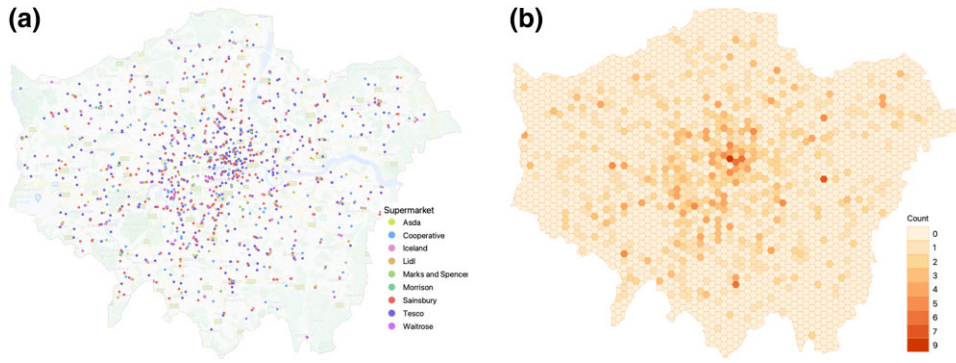

Figure 10. (a) Visualisation of the supermarket locations with their respective supermarket chains name. (b) Greater London is split into equal size grids of hexagons (size of each side is 0.5 km) and number of supermarkets within each hexagon is displayed.

ownership data with the National polygons dataset (HM Land Registry, 2020a) to identify each properties title polygon. Finally, we spatially join the two datasets to obtain a dataset of Supermarkets in the UK with their geo-location coordinates. The flow diagram of the process is demonstrated such that it is visually easy to read Figure 9. The spatial distribution of supermarkets in Greater London is shown in Figure 10.

In addition to the supermarket floor space, we calculate the Euclidean distance to the closest public transport access points (Department for Transport, 2014), tourist attractions (Historic England, 2014) to describe the store characteristics. The customer rating given for each supermarket store is accessed using the Google place API (Google, 2020). We extract the annual revenue generated by supermarkets using the annual statements published by the companies. Since the individual revenues at each store are not published, we calculate a revenue proxy using the reported annual revenues.

5 Case studies: optimal locations for supermarkets and pubs in London

In this section, we illustrate our proposed optimisation algorithm with the hierarchical search using two real-world datasets. Initially, we demonstrate the BSIM by modelling the revenue of supermarkets and subsequently find the optimal locations for supermarkets and pubs in Greater London.

5.1 The supermarkets in London

The first case study is centred on the seven largest supermarket chains in the UK. We construct a complete dataset for $S = 1,079$ supermarkets located within Greater London. We use the derived store features 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.

5.1.1 Estimating revenues using the BSIM

We estimate the BSIM parameters under four truncated radii for the Gaussian distribution and summarise the performance in Table 4 with standard metrics.

(a) 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 ($\mathbf{Y}$), $\text{NRMSE} = \sqrt{E[\mathbf{Y} - \hat{\mathbf{Y}}]^2}/E[\mathbf{Y}]$.

Table 4. R^2, σ^2, and NRMSE for the fitted BSIM for revenues of supermarkets in London under four different radii of the truncated Gaussian distribution

	Truncated radius (km)			
Performance metric	10	15	20	25
R^2	0.38	0.64	**0.89**	0.1
σ^2	0.64	0.5	0.31	0.72
NRMSE	0.07	0.05	0.03	0.09

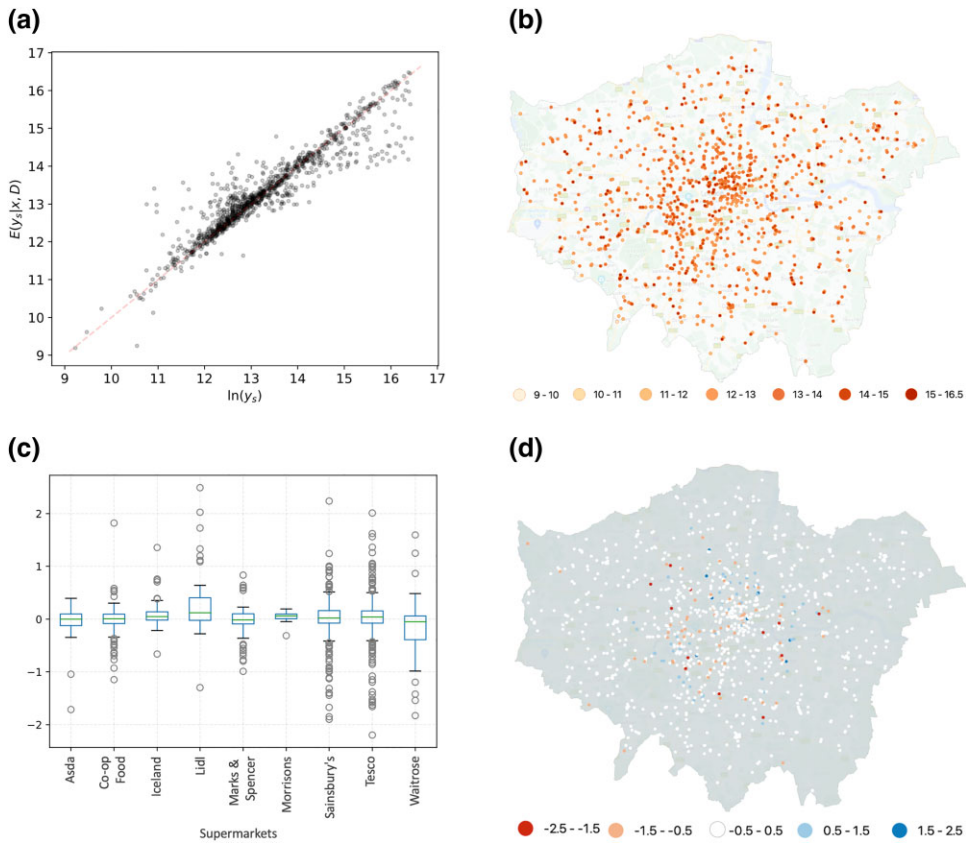

Figure 11. The results of the best-performed experiment for the BSIM with the supermarkets' revenue. (a) Actual against predicted revenue. (b) Predicted revenue at each store. (c) Residuals against the supermarket chain. (d) Spatial distribution of the residuals.

(b) 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}), $R^2 = 1 - SS_{res}/SS_{tot}$.

We assume the reasonable distance a customer is willing to travel is half the maximum extent that 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 decrease significantly as it reaches a 25 km radius that covers the whole of London. The results from the best-fitted model are demonstrated in Figure 11. The scatter plot with the actual log revenue against the predicted log revenue in Figure 11a shows

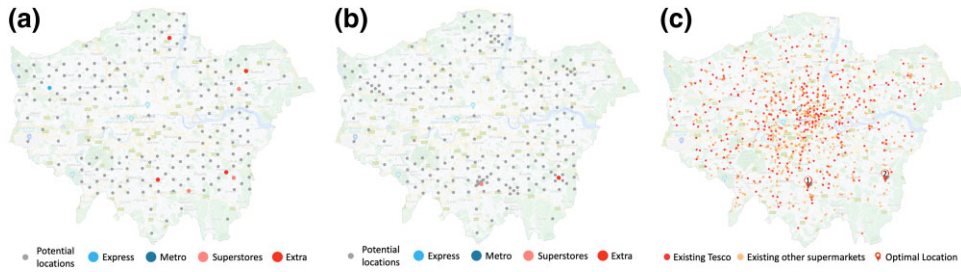

Figure 12. Optimal locations to establish two Tesco supermarkets. (a) The initial set of candidate locations is generated from multiresolution sampling with 5×5 grids with a depth of three and optimal locations from four independent samples. (b) All the potential locations that were evaluated at different stages and the final optimal locations. (c) Existing Tesco and other supermarkets and new optimal stores.

Table 5. Monthly revenue estimations of the optimal stores reported in millions

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

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 Figure 11b. We explore the residual values for each supermarket chain in Figure 11c. 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 Figure 11d.

5.1.2 Optimal location

We use the parameter estimates from the best-fitted BSIM to calculate the objective function of the optimisation problem. There are four types of supermarket stores with varying floorspace: Express (278 sqm), Metro (1,021 sqm), Superstores (3,251 sqm), and Extra (5,574 sqm). We use these sizes as the possible designs to structure the new facilities. Additionally, we calculate the other characteristics at each potential facility, and for Google ratings, it is assumed to have the average ratings of the existing stores for each chain. 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. We search for optimal locations to build at most two supermarkets using a budget of 35. The optimal locations are demonstrated for the largest supermarket chain in UK, Tesco. 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, we use equation 18 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 Figure 12a. The search algorithm continued for two iterations evaluating the neighbourhood locations produced from quadtree. The optimal locations for Tesco supermarket are detected to be in Croydon and Bromley with designs of a Superstore and Extra, respectively, as shown in Figure 12b. No new facility is to be located in the same area as one of the competitors. The median and 50% credible interval (CI) of the estimated revenue for the two optimal supermarkets are reported in Table 5. Both the facilities are predicted to generate more revenue than the average revenue produced by the existing supermarkets in their respective Boroughs. The

Table 6. Characteristics of the two optimal supermarkets

			Stores in 5 km radius †		Distance to the nearest (m)		
Supermarket	Design	Floor size (sqm)	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.

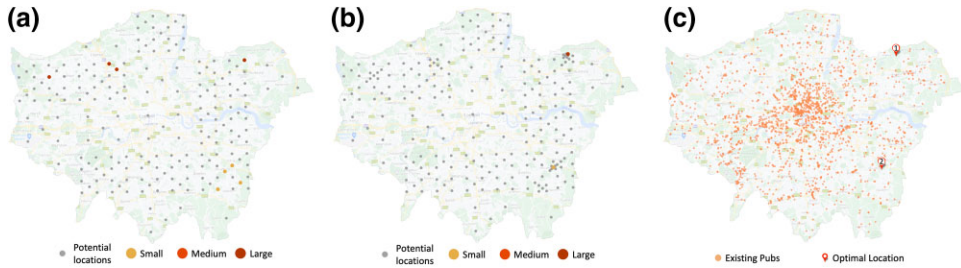

Figure 13. Optimal sites to establish two pubs in London. (a) Optimal locations from four independent samples. (b) All the potential locations that were eventuated at different stages and the final optimal locations. (c) Existing pubs and new optimal facility location.

Table 7. Monthly revenue† estimations of the two optimal pubs reported in millions

			Estimated Revenue	
Pub	Borough	Average revenue in the Borough	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 8. Characteristics of the two optimal pubs

			Distance to the nearest (m)					
Pub	Design	Floor size (sqm)	Metro	Rail	Bus	Parks	Attractions	Sports facility
1	Small	175	836	2,136	533	312	224	678
2	Large	1,275	3,521	4,564	733	6,206	4,285	718

recommended new supermarket in Bromley is located in a less dense area as shown in Figure 12c. There is only one Tesco and 16 other supermarkets in a 5 km radius (Table 6), 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.

The experiment is extended to find the optimal location with a budget of 40 that would assess all possible combinations of supermarket types. The optimal locations remain the same as the

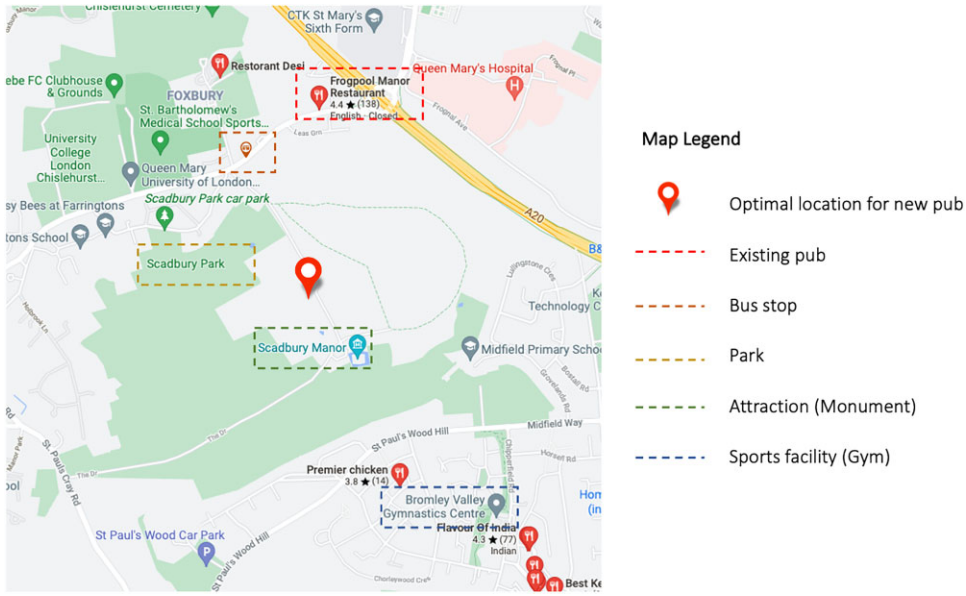

Figure 14. Eagle view of the optimal pub location with the small design. The dashed squares indicate some of the key venues in the surrounding of 1 km radius.

previous setting, but both supermarkets are recommended to be of Extra type. The estimated revenue of the store in Croydon is marginally increased to 3.2 million (2.8, 4.3)[1] .

5.2 Optimal locations to enter the pubs market in London

In our next real-world case study, we use the data and parameter estimates evaluated in the spatial interactions study by Perera et al. (2021). Optimisation problem considers three sizes of pub designs with total floor size: 175 sqm, 500 sqm, and 1,275 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. We search for optimal locations to build at most two pubs using a budget of nine. The same sampling approach is used to generate the initial set of potential locations as described for the supermarket experiment. 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, Figure 13b.

The median and 50% credible interval (CI) of the estimated revenue for the two locations are displayed in Table 7. The monthly estimated sales of both the pubs are higher than the average revenue generated by the existing pubs within their respective boroughs. Distance between the optimal locations and the public transport access points and key venues are presented in Table 8. New sites are located near sports facilities and closer to bus stops. Revenue at the small pub is expected to be driven by the customers attracted to the area with key venues.

The experiment is extended to find the optimal location with a budget of 14 that would assess all possible combinations of designs. The two optimal sites are detected in Bromley with large-sized pubs. The pubs are estimated to generate monthly revenue of 3.7 million (2.2, 8.8) and 3.6 million (2.2, 7.9). We explore the area of the facility for a small pub identified in the first experiment with an eagle view in Figure 14. 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 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.

[1] 50% CI of estimated revenue

6 Discussion

We have formulated a mathematical modelling framework to simultaneously identify optimal facility locations and corresponding designs in a competitive environment. This formulation considerably improves the existing competitive models based on classical utility models as it considers model uncertainty via a Bayesian approach and provides probability density estimates of the revenue at new stores. In estimating the revenue, we extend the recently introduced BSIM by lifting the assumption of fixed demand by introducing dummy facilities to make more realistic estimations.

We 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 under-represent the true combinations of optimal locations when searching for more than one facility. The initial candidate locations created with the multi-resolution grid structure that accounts for density between customer spending and existing facilities reported the best and most efficient results. The optimisation framework and the applications in this study assume that the company owns the chain of facilities, and the objective is to maximise the total revenue. However, in the case of a franchise system, cannibalisation of existing chain outlets is minimised so as not to gain market share at the expense of member outlets (Drezner, 2011; Pelegrín et al., 2016). The proposed optimisation framework can be modified by adding a constraint to account for cannibalisation.

Unique to this paper, we present a fully integrated large-scale, real-world application by first estimating the spatial interactions and subsequently locating the optimal sites for the largest supermarket chain in UK to expand their 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 we present our methodology for supermarkets and pubs, it 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 for the cost may produce more realistic results.

The proposed methodology can be extended and improved in future research in multiple directions. Extending the framework to include temporal dynamics could provide better recommendations to place the new facilities by accounting for the changes in the urban systems. Also, one could extend the framework to deal with uncertainty in the data of the optimisation problem by applying robust optimisation (Berger et al., 1994).The sampling method for generating potential locations can be advanced by applying determinantal point process (Loonis & Mary, 2019) where the kernel matrix is computed using the store locations to capture the competition among the existing stores and combine with self-exciting point process (Mohler et al., 2011) to cluster stores where the spending capacity of the population is higher. Furthermore, considering industry-led cost functions for placement or risk exposure is an interesting extension of our work that could be studied under Bayesian decision theory (Berger, 2013).

Acknowledgments

We would like to thank the UK Engineering and Physical Sciences Research Council (EPSRC grant no. EP/L016710/1 and EP/R512229/1). Furthermore, this work was supported by the Alan Turing Institute under EPSRC grant EP/N510129/1, UKRI Turing AI fellowship EP/V02678X/1, and the Lloyds Register Foundation. We are also grateful to Nimbus property system limited for their support and giving access to a comprehensive property database, and for the valuable insights shared by its Director, Paul Davis.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

Conflict of interests: None declared.

Data availability

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.

References

Benati S., & Hansen P. (2002). The maximum capture problem with random utilities: Problem formulation and algorithms. *European Journal of Operational Research*, *143*(3), 518–530. https://doi.org/10.1016/S0377-2217(01)00340-X

Berger J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Berger J. O., Moreno E., Pericchi L. R., Bayarri M. J., Bernardo J. M., Cano J. A., De la Horra J., Martín J., Ríos-Insúa D., Betrò B., Dasgupta A., Gustafson P., Wasserman L., Kadane J. B., Srinivasan C., Lavine M., O'Hagan A., Polasek W., Robert C. P., Sivaganesan S. (1994). An overview of robust Bayesian analysis. *Test*, *3*(1), 5–124. https://doi.org/10.1007/BF02562676

Berman O., Drezner T., Drezner Z., & Krass D. (2009). Modeling competitive facility location problems: New approaches and results. In *Decision technologies and applications* (pp. 156–181). INFORMS.

Bishop C. (2006). *Pattern recognition and machine learning*. Information Science and Statistics. Springer.

Cressie N. (1994). 4-models for spatial processes. In J. L. Stanford & S. B. Vardeman (Eds.), *Statistical methods for physical science*, vol. 28 of *Methods in experimental physics* (pp. 93–124). Academic Press.

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

Drezner T. (2006). Derived attractiveness of shopping malls. *IMA Journal of Management Mathematics*, *17*(4), 349–358. https://doi.org/10.1093/imaman/dpl004

Drezner T. (2011). Cannibalization in a competitive environment. *International Regional Science Review*, *34*(3), 306–322. https://doi.org/10.1177/0160017610389328

Drezner T. (2014). A review of competitive facility location in the plane. *Logistics Research*, *7*(1), 114. https://doi.org/10.1007/s12159-014-0114-z

Drezner T., & Drezner Z. (2012). Modelling lost demand in competitive facility location. *Journal of the Operational Research Society*, *63*(2), 201–206. https://doi.org/10.1057/jors.2011.10

Drezner T., Drezner Z., & Kalczynski P. (2012). Strategic competitive location: Improving existing and establishing new facilities. *Journal of the Operational Research Society*, *63*(12), 1720–1730. https://doi.org/10.1057/jors.2012.16

Freire A. S., Moreno E., & Yushimito W. F. (2016). A branch-and-bound algorithm for the maximum capture problem with random utilities. *European Journal of Operational Research*, *252*(1), 204–212. https://doi.org/10.1016/j.ejor.2015.12.026

Gisbert F. J. G. (2003). Weighted samples, kernel density estimators and convergence. *Empirical Economics*, *28*(2), 335–351. https://doi.org/10.1007/s001810200134

Google (2020). Place search. https://developers.google.com/places/web-service/search.

Harold H. (1929). Stability in competition. *Economic Journal*, *39*(153), 41–57.

Historic England (2014). Listing. https://historicengland.org.uk/listing/the-list/.

HM Land Registry (2020a). National polygon service. https://www.gov.uk/guidance/national-polygon-service.

HM Land Registry (2020b). UK companies that own property in England and Wales. https://www.gov.uk/guidance/hm-land-registry-uk-companies-that-own-property-in-england-and-wales.

Huff D. L. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics*, *39*(1), 81. https://doi.org/10.2307/3144521

Jordan M. I., Ghahramani Z., Jaakkola T. S., & Saul L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233. https://doi.org/10.1023/A:1007665907178

Kellerer H., Pferschy U., & Pisinger D. (2004). The multiple-choice knapsack problem. In *Knapsack problems* (pp. 317–347). Springer.

Küçükaydin H., Aras N., & Altınel I. K. (2011). 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. https://doi.org/10.1016/j.ejor.2010.08.009

Leonardi G., & Tadei R. (1984). Random utility demand models and service location. *Regional Science and Urban Economics*, *14*(3), 399–431. https://doi.org/10.1016/0166-0462(84)90009-7

Lewis P. W., & Shedler G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, *26*(3), 403–413. https://doi.org/10.1002/nav.3800260304

Loonis V., & Mary X. (2019). Determinantal sampling designs. *Journal of Statistical Planning and Inference*, *199*(1), 60–88. https://doi.org/10.1016/j.jspi.2018.05.005

Mohler G. O., Short M. B., Brantingham P. J., Schoenberg F. P., & Tita G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546

Murray A. T. (2018). Evolving location analytics for service coverage modeling. *Geographical Analysis*, *50*(3), 207–222. https://doi.org/10.1111/gean.12146

ONS (2020). Retail sales index internet sales. https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/retailsalesindexinternetsales.

Ordnance Survey (2019). AddressBase Premium. https://www.ordnancesurvey.co.uk/business-government/products/addressbase-premium.

Pelegrín B., Fernández P., & Garcia Perez M. D. (2016). Profit maximization and reduction of the cannibalization effect in chain expansion. *Annals of Operations Research*, *246*(1–2), 57–75. https://doi.org/10.1007/s10479-014-1676-5

Perera S., Aglietti V., & Damoulas T. (2021). 'A variational Bayesian spatial interaction model for estimating revenue and demand at business facilities', arXiv, arXiv:2108.02594, preprint: not peer reviewed.

Reilly W. J. (1931). *The law of retail gravitation*. WJ Reilly.

Samet H. (2006). *Foundations of multidimensional and metric data structures*. Morgan Kaufmann.

Silverman B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability, vol. 26. Chapman and Hall.

Toregas C., Swain R., ReVelle C., & Bergman L. (1971). The location of emergency service facilities. *Operations Research*, *19*(6), 1363–1373. https://doi.org/10.1287/opre.19.6.1363

Valuation Office Agency (2019). Business rates. https://www.gov.uk/introduction-to-business-rates.

van de Schoot R., Depaoli S., King R., Kramer B., Märtens K., Tadesse M. G., Vannucci M., Gelman A., Veen D., Willemsen J., & Yau C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1–26. https://doi.org/10.1038/s43586-020-00001-2

Wilson A. G. (1971). A family of spatial interaction models, and associated developments. *Environment and Planning A*, *3*(1), 1–32. https://doi.org/10.1068/a030001

Wolsey L. A., & Nemhauser G. L. (1999). *Integer and combinatorial optimization*, vol. *55*. John Wiley & Sons.