New continuous approximation models for passenger and freight transportation

July 2022

A Research Report from the Pacific Southwest Region University Transportation Center

John Gunnar Carlsson, University of Southern California

PSR Pacific Southwest Region UTC
University Transportation Center

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. PSR-21-22	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle New continuous approximation models for passenger and freight transportation		**5. Report Date** 07/29/2022
		6. Performing Organization Code N/A
7. Author(s) John Gunnar Carlsson, 0000-0001-5346-8529		**8. Performing Organization Report No.** TBD
9. Performing Organization Name and Address METRANS Transportation Center University of Southern California University Park Campus, RGL 216 Los Angeles, CA 90089-0626		**10. Work Unit No.** N/A **11. Contract or Grant No.** USDOT Grant 69A3551747109
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final report (08/16/2021-08/15/2022) **14. Sponsoring Agency Code** USDOT OST-R
15. Supplementary Notes		

16. Abstract

The purpose of this project is to discover new *continuous approximation models* for modern logistical problems, such as last-mile delivery and the adoption of teleworking. The continuous approximation paradigm is a quantitative method for solving logistics problems in which one uses a small set of parameters to model a complex system, which results in a simple algebraic equation that is easier to manage than (for example) a large-scale optimization model. As a further benefit, one often obtains insights from these simpler formulations that help to determine what affects the outcome most significantly. Continuous approximation models have been used for over 60 years to study classical logistical problems, but modern logistical systems bring new levels of complexity that existing models do not address. This project combines tools from geospatial optimization, computational geometry, and geometric probability theory to formulate new models that will enable practitioners and policy-makers to solve these new problems, and most importantly, to identify what features are most impactful in their real-world use.

17. Key Words Vehicle routing; trip chaining; last mile delivery	**18. Distribution Statement** No restrictions.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 39	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Contents

About the Pacific Southwest Region University Transportation Center

The Pacific Southwest Region University Transportation Center (UTC) is the Region 9 University Transportation Center funded under the US Department of Transportation's University Transportation Centers Program. Established in 2016, the Pacific Southwest Region UTC (PSR) is led by the University of Southern California and includes seven partners: Long Beach State University; University of California, Davis; University of California, Irvine; University of California, Los Angeles; University of Hawaii; Northern Arizona University; Pima Community College.

The Pacific Southwest Region UTC conducts an integrated, multidisciplinary program of research, education and technology transfer aimed at *improving the mobility of people and goods throughout the region.* Our program is organized around four themes: 1) technology to address transportation problems and improve mobility; 2) improving mobility for vulnerable populations; 3) Improving resilience and protecting the environment; and 4) managing mobility in high growth areas.

U.S. Department of Transportation (USDOT) Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Disclosure

John Gunnar Carlsson conducted this research titled, "New continuous approximation models for passenger and freight transportation" at the Epstein Department of Industrial and Systems Engineering at the Viterbi School of Engineering at the University of Southern California. The research took place from August 2021 to August 2022 and was funded by a grant from METRANS in the amount of $99,998. The research was conducted as part of the Pacific Southwest Region University Transportation Center research program.

Acknowledgements

The authors gratefully acknowledge the support of METRANS for funding this research.

Abstract

The purpose of this project is to discover new *continuous approximation models* for modern logistical problems, such as last-mile delivery and the adoption of teleworking. The continuous approximation paradigm is a quantitative method for solving logistics problems in which one uses a small set of parameters to model a complex system, which results in a simple algebraic equation that is easier to manage than (for example) a large-scale optimization model. As a further benefit, one often obtains insights from these simpler formulations that help to determine what affects the outcome most significantly. Continuous approximation models have been used for over 60 years to study classical logistical problems, but modern logistical systems bring new levels of complexity that existing models do not address. This project combines tools from geospatial optimization, computational geometry, and geometric probability theory to formulate new models that will enable practitioners and policy-makers to solve these new problems, and most importantly, to identify what features are most impactful in their real-world use.

New continuous approximation models for passenger and freight transportation

Executive Summary

The purpose of this project is to design simple and concise mathematical models for quantifying the reductions in vehicle miles travelled (VMT), among other cost measures, that result from implementing modern logistics systems with complex features such as ridesharing, crowdsourcing or teleworking. Traditionally, these problems have been solved in a discrete setting, involving fixed sets of (for example) demand points, time periods, and service facility locations; one then solves them with an integer mathematical programming solver such as CPLEX or Gurobi. A drawback of this approach is that the problems are almost always *NP-hard*, and hence solving large-scale instances would require enormous computational efforts which likely increase exponentially with the problem instance size. A further drawback is that such models are often extremely complex, which hinders understanding of salient problem features and managerial insights.

For these reasons, this project has used tools from geospatial optimization, computational geometry, and geometric probability theory to discover simple *continuous approximation models* that identify the key problem attributes that affect them most significantly. A continuous approximation model is characterized by its use of continuous representations of input data and decision variables as density functions over time and space, and the goal is to approximate the objective function into an expression that can be optimized by relatively simple analytical operations. Such an approximation enables transforming otherwise high-dimensional decision variables into a low-dimensional space, allowing the optimal solution to be obtained with mere calculus, even when significant operational complexities are present. The results from such models often bear closed-form analytical structures that help reveal managerial insights.

Continuous approximation models for classical transportation problems such as the travelling salesman problem (TSP) and the vehicle routing problem (VRP) are well-understood. Modern logistical challenges, such as the integration of crowdsourcing, ridesharing, and so-called "random stow" policies, all bring with them new problem complexities that are not easily handled using traditional continuous approximation models. However, our recent successes in previous research sponsored by METRANS indicate to us that there do indeed exist tractable continuous approximation models for handling them, and we present them in this report.

1 Introduction

Recent years have seen major changes in modern logistics systems analysis, due to the adoption of modalities such as ridesharing, crowdsourcing or teleworking. New paradigms such as these often give rise to challenging mathematical optimization problems, such as mixed-integer linear programs (MILPs) or constraint programs. Traditionally, these problems have been solved in a discrete setting, involving fixed sets of (for example) demand points, time periods, and service facility locations; one then solves them with an integer mathematical programming solver such as CPLEX or Gurobi. A drawback of this approach is that the problems are almost always NP-hard, and hence solving large-scale instances would require enormous computational efforts which likely increase exponentially with the problem instance size. A further drawback is that such models are often extremely complex, which hinders understanding of salient problem features and managerial insights.

For these reasons, this project has used tools from geospatial optimization, computational geometry, and geometric probability theory to discover simple *continuous approximation models* that identify the key problem attributes that affect them most significantly. A continuous approximation model is characterized by its use of continuous representations of input data and decision variables as density functions over time and space, and the goal is to approximate the objective function into an expression that can be optimized by relatively simple analytical operations. Such an approximation enables transforming otherwise high-dimensional decision variables into a low-dimensional space, allowing the optimal solution to be obtained with mere calculus, even when significant operational complexities are present. The results from such models often bear closed-form analytical structures that help reveal managerial insights. Continuous approximation models for classical transportation problems such as the travelling salesman problem (TSP) and the vehicle routing problem (VRP) are well-understood. Modern logistical challenges, such ridesharing and so-called "random stow" policies in materials handling, all bring with them new problem complexities that are not easily handled using traditional continuous approximation models.

One family of problems that we found particularly relevant for this study are what we call *selection routing problems*. A selection routing problem is a routing optimization problem, such as the TSP or VRP, in which one is given a large collection of destinations and the goal is to select a subset of those points that satisfies certain criteria and optimizes some objective function. By way of comparison, the TSP and VRP both require that one visit all destinations. Such problems are particularly timely in modern analysis of logistical systems in several contexts:

- Selection routing problems arise organically in studying the consequences of *trip chaining* [24], that is, performing multiple errands during a single outing, because one has multiple choices of locations at which to perform errands.

- One proposed approach for mitigating the inefficiencies in "last mile" delivery has been the use of a socially networked system in which parcel recipients can "opt in" for packages to be delivered at *multiple* possible locations (as opposed to their doorstep), such as their workplace [27, 38]. The parcel delivery company then solves a selection routing problem in which they must select one of the multiple locations for each customer and deliver a package there.

- Selection routing problems are fundamentally important in studying randomized strategies

in warehouses, in which one stores a stock keeping unit (SKU) in any available location (as opposed to designating specific regions of the warehouse for different SKUs). This is because a warehouse picker will often select multiple SKUs at a time, and can benefit if those SKUs are dispersed throughout the warehouse. Amazon, for example, calls this process *random stow* and attributes its rapid growth to the efficiency that is realized as a result [8]:

> **Random stow:** The storage of items in a randomised order at fulfilment centres to maximise the chance of multiple items on the same order being near each other. The fulfilment centre management system knows the location of every item and is able to work out the shortest travel distance to pick the orders.

The remainder of this report is devoted to the derivation of new continuous approximation models for various selection routing problems, which we subsequently verify and validate using computational experiments.

2 Related research

This section reviews previous research in the area of selection routing problems as well as continuous approximation models in logistics.

2.1 Selection routing problems

One of the most famous selection routing problems is the *generalized TSP* (GTSP). In the GTSP, given a collection of point sets, $\mathscr{X}_1, \ldots, \mathscr{X}_n$, we seek the shortest tour that visits 1 point of each point set $\mathscr{X}_i$; the traditional TSP simply has $\mathscr{X}_i = \{x_i\}$ for all i, i.e. each point set is a single point. The GTSP was first introduced in [1], and the author proposes a dynamic programming approach for solving it. The application in their paper is sequencing computer files. The GTSP is an NP-Hard optimization problem with potential applications in warehousing, distribution, and scheduling. As such it is important to solve it to a high level of optimality. Saksena [32] studied the routing of welfare clients through governmental agencies as a symmetric GTSP. Other applications include vehicle dispatching [14], plant location [30], and other problems such as the warehouse order picking with multiple stock locations, airport selection and routing for courier planes, and certain types of flexible manufacturing scheduling. For more on these applications, we invinte the reader to refer to papers such as [12] and [23].

2.2 Continuous approximation models

The motivation for studies on continuous approximation paradigm is the replacement of combinatorial quantities that are difficult to compute with simpler mathematical expressions, which (under certain conditions) provide accurate estimations. This type of approximation is common for combinatorial problems. Examples include TSP approximations as in [2], which is reproduced in Theorem 4 of this report, and [10], facility location problems as in [17], [19], and [26], and basically any *subadditive Euclidean functional* such as a minimum spanning tree [35], k-medians [20], Steiner tree [35], or matching [33] and other papers such as [28], [34], and [36] papers.

This report derives continuous approximation models for selection routing problems. In this way, it is similar to papers like [4], which studies the trade-offs between inventory and transportation costs. Also, the work in [21] studies how to route relief vehicles following a disaster in a time-sensitive manner. Chowdhury et al. [7] showed how to optimally partition the disaster-affected region for emergency drone distribution. Another similar work is [22] which develops a geometric model to find the optimal long-term vehicle fleet composition for distribution activities. Finally for more refere to Franceschetti et al. who reviewed many studies in the same vein in their paper [13].

In this work, we assume that the points are distributed based on a density function, i.e. stochastically. The paper [5] studies a partitioning algorithm in which the client locations are independent and identically distributed samples from a given probability density function. Goodson [15] describes a set of rollout policies based on fixed routes to obtain dynamic solutions to the vehicle routing problem with stochastic demand and duration limits. For more on the literature of stochastic vehicle routing models refer to the survey in [31].

3 Preliminaries

The first four results are stated without proof and are standard textbook material. Theorem 4 is the famous *Beardwood-Halton-Hammersley (BHH) Theorem* [2].

Lemma 1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function and let $\mathfrak{B}_d(r) \subset \mathbb{R}^d$ be a ball of radius r centered about the origin. We have*

$$\int_{\mathfrak{B}_d(r)} f(\|\mathbf{x}\|)\, d\mathbf{x} = \int_0^r S_{d-1}(t) f(t)\, dt\,,$$

where $S_{d-1}(t)$ is the surface area of a $(d-1)$-sphere of radius t, which is given by

$$S_{d-1}(t) = \frac{2\pi^{d/2}}{\Gamma(d/2)} t^{d-1}\,.$$

Lemma 2. *The volume of a d-dimensional ball of radius r is $\pi^{d/2} r^d / \Gamma(d/2+1)$, where Γ denotes the gamma function..*

Lemma 3 (Stirling's formula)**.** *The gamma function $\Gamma(x)$ satisfies* $\log \Gamma(x+1) = x \log x - x + \frac{1}{2}\log x + \frac{1}{2}\log 2 + \frac{1}{2}\log \pi + \mathcal{O}(1/x)$ *as $x \to \infty$.*

Theorem 4 (BHH Theorem)**.** *There is a constant β_d such that, for almost any sequence of independent random variables $\{X_i\}$ sampled from an absolutely continuous density f on $\mathbb{R}^d$ with compact support, we have*

$$\lim_{n\to\infty} \frac{\mathrm{TSP}(X_1, \ldots, X_n)}{n^{(d-1)/d}} = \beta_d \int_{\mathbb{R}^d} f(\mathbf{x})^{(d-1)/d}\, d\mathbf{x}$$

with probability one.

Lemma 5 (Super- and sub-additivity of the TSP)**.** *Let $\mathcal{R} \subset \mathbb{R}^2$ be a compact Lebesgue measurable set, partitioned into pieces $\mathcal{P}_1, \ldots, \mathcal{P}_m$ whose common boundaries (i.e. $\partial\mathcal{P}_i \cap \partial\mathcal{P}_j$) have finite*

length. There exists a constant C that depends only on the partition such that, for any set of points $X = \{x_1, \ldots, x_n\} \subset \mathcal{R}$, we have

$$-C + \sum_{i=1}^{m} \mathsf{TSP}(X \cap \mathcal{P}_i) \leq \mathsf{TSP}(X) \leq C + \sum_{i=1}^{m} \mathsf{TSP}(X \cap \mathcal{P}_i)$$

Proof. See Lemma 2.4.1 of [36], for example. □

Lemma 6 (Borel-Cantelli)**.** *Let $\{E_n\}$ be a sequence of events in a sample space. Then if $\sum_{n=1}^{\infty} \Pr(E_n) < \infty$, we have*

$$\Pr(E_n \text{ occurs infinitely often}) = 0\,,$$

or equivalently

$$\Pr(\limsup_{n\to\infty} E_n) := \Pr(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m) = 0\,.$$

The remaining results are routine volume computations:

Lemma 7. *Let $\ell > 0$ and let $\mathcal{D} \subset \mathbb{R}^{dn}$ denote the set of all n-tuples $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ of points in $\mathbb{R}^d$ such that $\sum_{i=1}^{n} \|\mathbf{u}_i\| \leq \ell$. The volume of $\mathcal{D}$, $\mathsf{Vol}(\mathcal{D})$, satisfies*

$$\mathsf{Vol}(\mathcal{D}) = \left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^n \cdot \frac{\Gamma(d)^n}{\Gamma(dn+1)} \cdot \ell^{dn}\,. \tag{1}$$

Proof. This is nothing more than the integral

$$\int_{\mathcal{B}_d(\ell)} \int_{\mathcal{B}_d(\ell - \|\mathbf{u}_n\|)} \cdots \int_{\mathcal{B}_d(\ell - \sum_{i=3}^{n} \|\mathbf{u}_i\|)} \int_{\mathcal{B}_d(\ell - \sum_{i=2}^{n} \|\mathbf{u}_i\|)} 1 \, d\mathbf{u}_1 \, d\mathbf{u}_2 \cdots d\mathbf{u}_{n-1} \, d\mathbf{u}_n\,,$$

which we can compute using a standard inductive argument applying Lemma 1. □

Corollary 8. *Let $\ell > 0$ and let $\mathcal{D}' \subset \mathbb{R}^{dn}$ denote the set of all n-tuples $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of points in $\mathbb{R}^d$ such that $\|x_1\| + \sum_{i=2}^{n} \|x_i - x_{i-1}\| \leq \ell$. The volume of $\mathcal{D}'$, $\mathsf{Vol}(\mathcal{D}')$, satisfies*

$$\mathsf{Vol}(\mathcal{D}') = \left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^n \cdot \frac{\Gamma(d)^n}{\Gamma(dn+1)} \cdot \ell^{dn}\,. \tag{2}$$

Proof. This is just Cavalieri's principle applied to Lemma 7 □

4 Two additional lemmas

The lemma below is useful for bounding the length of a TSP tour from below:

Lemma 9. *Let X_0 be the origin in $\mathbb{R}^d$ and let $X_1, \ldots, X_n$ be a collection of independent, uniform samples drawn from a region $\mathcal{R}$ of unit volume in $\mathbb{R}^d$. Then*

$$\Pr(\mathsf{TSP}(X_0, X_1, \ldots, X_n) \leq \ell) \leq \Gamma(n+1) \cdot \left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^n \cdot \frac{\Gamma(d)^n}{\Gamma(dn+1)} \cdot \ell^{dn}\,.$$

Proof. It is easy to see that

$$\underbrace{\text{Pr}\quad \|X_1\| + \sum_{i=2}^{n} \|X_i - X_{i-1}\| \le \ell}_{(*)} \le \left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^n \cdot \frac{\Gamma(d)^n}{\Gamma(dn+1)} \cdot \ell^{dn}; \tag{3}$$

this is because we can regard the samples $X_1, \ldots, X_n$ as being a *single* sample drawn uniformly from $\mathcal{R}^n$, so that the probability of interest $(*)$ is simply the probability that this single sample lies in the domain $\mathcal{D}'$ described in Corollary 8. This probability is of course equal to $\text{Vol}(\mathcal{D}' \cap \mathcal{R}^n) \le \text{Vol}(\mathcal{D}')$, which gives us the desired inequality (3). We obtain our lemma by applying the union bound to (3) over all $n! = \Gamma(n+1)$ permutations of $X_1, \ldots, X_n$. □

The following lemma allows us to approximate a probability density function with a step function in a way that preserves certain quantities of interest.

Lemma 10 (Approximation with a simple function)**.** *Let f be a probability density function with compact support $\mathcal{R} \subset \mathbb{R}^2$ whose level sets have Lebesgue measure zero, let k be a positive integer, and define*

$$P(x) = \text{Pr}(f(X) \le f(x)) = \iint_{x':f(x') \le f(x)} f(x')\ \mathrm{d}x'\,.$$

For any $\epsilon > 0$, there exists a step density function $\phi(x) = \sum_{i=1}^{s} a_i \mathbb{1}(x \in \boxdot_i)$ and corresponding

$$\Pi(x) = \text{Pr}(\phi(X) \le \phi(x)) = \iint_{x':\phi(x') \le \phi(x)} \phi(x')\ \mathrm{d}x'$$

such that the following conditions hold:

1. $\int_{\mathcal{R}} |\phi(x) - f(x)|\ \mathrm{d}x \le \epsilon$,
2. $\left|P(x)^{k-1} - \Pi(x)^{k-1}\right| \le \epsilon$ for all $x \in \mathcal{R}$,
3. All of the components of ϕ have the same mass, i.e. $a_i\,\text{Area}(\boxdot_i) = 1/s$.

Proof. The requirement that the level sets have measure zero just is not actually necessary, but we find it useful for keeping notation consistent throughout this paper (this requirement is flagrantly violated when f is a uniform distribution, which is our base case for all of the various TSP instances in this paper anyway). For a large integer q, define contour sets $\mathcal{S}_i = \{x : (i-1)/q < P(x) \le i/q\}$. For each $\mathcal{S}_i$, we can approximate the restriction of f to $\mathcal{S}_i$ (i.e. $f(x)\mathbb{1}(x \in \mathcal{S}_i)$) to arbitrary precision ϵ' by a step function $\psi_i(x) = \sum_j a_{ij} \mathbb{1}\{x \in \boxdot_{ij}\}$ (this is a classical result of measure theory; see e.g. Theorem 2.4(ii) of [37]). Section A of the appendix also establishes that for all i and j, we can assume without loss of generality that

- $\boxdot_{ij} \subset \mathcal{S}_i$ for all i and j (i.e. the support of ψ_i is contained in $\mathcal{S}_i$),
- $a_{ij} < a_{(i+1)j'}$ for all i, j, and j',
- all a_{ij} and $\text{Area}(\boxdot_{ij})$ are rational, and

- $\int_{\mathcal{S}_i} \psi_i = \int_{\mathcal{S}_i} f = 1/q$.

Given any $\epsilon > 0$, we set $q = \lceil 1/\epsilon \rceil$ and $\epsilon' = \epsilon/q$ in the above construction. The function $\psi := \sum_{i=1}^{q} \psi_i$ is therefore a step density approximation of f whose aggregate error over $\mathcal{R}$ is at most ϵ, so condition 1 is satisfied. If we define $\Pi'(x) = \int_{x':\psi(x')\le\psi(x)} \psi(x)\ \mathrm{d}x$, then condition 2 is satisfied as well; indeed, this is the purpose of the initial decomposition of $\mathcal{R}$ into the $\mathcal{S}_i$'s.

For ease of notation, we now re-index all of the components of ψ (i.e. we disregard the fact that ψ decomposes into a sum of ψ_i's) so that we simply have $\psi(x) = \sum_j b_j \mathbb{1}\{x \in \boxdot_j\}$, where b_j and $\mathrm{Area}(\boxdot_j)$ are rational. If we take δ to be the lowest common denominator over all b_j $\mathrm{Area}(\boxdot_j)$, then we can write $b_j\ \mathrm{Area}(\boxdot_j) = z_j/\delta$, with z_j a positive integer. To satisfy the third condition, all that remains is to decompose each $\boxdot_j$ into z_j pieces of equal area, and let ϕ denote the step function resulting thereof, which completes the proof. □

5 The main theorems proven in this report

The main results that we have proven in this report, as a result of this project, are shown below. We give their proofs in Sections 6 and 7. For both of these, we let f, $\mathcal{R}$, and P be as in Lemma 10.

Theorem (Section 6). *Let $\mathscr{X}_i \subset \mathbb{R}^2$ be a finite set of points, for $i \in \{1, \dots, n\}$, and assume that all sets $\mathscr{X}_i$ consist of k independent samples from f. Let $L(\mathscr{X}_1, \dots, \mathscr{X}_n)$ denote the length of the shortest tour that visits one element from each set $\mathscr{X}_i$. With probability one, we have*

$$\frac{\lambda_1}{2} < \liminf_{n\to\infty} \frac{L(\mathscr{X}_1, \dots, \mathscr{X}_n)}{\sqrt{nk} \iint_{\mathcal{R}} \sqrt{f(x)} P(x)^{k-1}\, \mathrm{d}x} \le \limsup_{n\to\infty} \frac{L(\mathscr{X}_1, \dots, \mathscr{X}_n)}{\sqrt{nk} \iint_{\mathcal{R}} \sqrt{f(x)} P(x)^{k-1}\, \mathrm{d}x} < \frac{\mu_1}{2} \qquad (4)$$

with $\lambda_1 = 0.4839$ and $\mu_1 = 1.8408$ as in Theorem 11.

Theorem (Section 7). *Let $X_1, \dots, X_n$ be a sequence of independent samples from f. Let $L(X_1, \dots, X_n; m)$ denote the length of the shortest tour that visits at m of the points $X_1, \dots, X_n$. For fixed $p \in (0, 1)$, we have*

$$\lambda_2 \int_{\mathcal{R}} \sqrt{f(x)} \mathbb{1}(P(x) \ge p)\, \mathrm{d}x < \liminf_{n\to\infty} \frac{L(X_1, \dots, X_n; \lceil pn \rceil)}{p\sqrt{n}} \le \limsup_{n\to\infty} \frac{L(X_1, \dots, X_n; \lceil pn \rceil)}{p\sqrt{n}} < \mu_2 \int_{\mathcal{R}} \sqrt{f(x)} \mathbb{1}(P(x) \ge p)\, \mathrm{d}x$$

with probability one, where $\lambda_2 = 0.2935$ and $\mu_2 = 0.9204$.

The value of the theorems above is that they provide a way to predict the length of a tour of a complex routing problem, without actually performing any optimization, by merely multiplying by $\sqrt{n}$. For instance, the second theorem suggests that a valid prediction is

$$L(X_1, \dots, X_n; pn) \approx cp\sqrt{n} \int_{\mathcal{R}} \sqrt{f(x)} \mathbb{1}(P(x) \ge p)\, \mathrm{d}x\,,$$

where c is a constant such that $\lambda_2 \le c \le \mu_2$. We will validate these claims in Section 8.

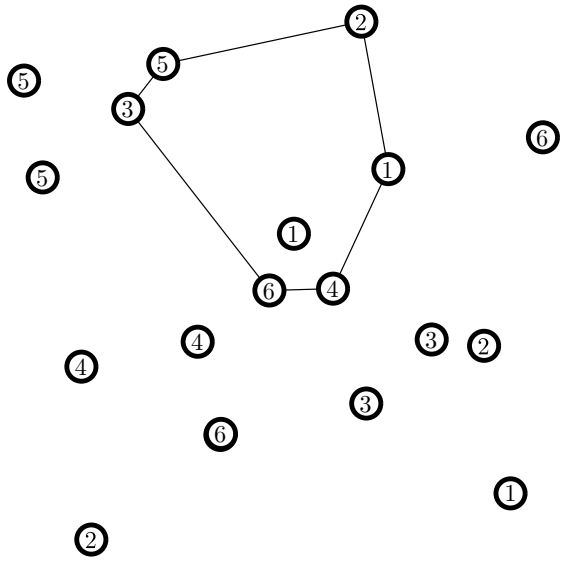


Figure 1: A generalized TSP tour of six sets of points $\mathscr{X}_1, \ldots, \mathscr{X}_6$, each consisting of $k = 3$ points. The optimal tour contains one element from each such set (and is the shortest such tour to do so).

6 Continuous approximation analysis of the generalized TSP (GTSP)

This section gives a continuous approximation formula for the GTSP; Theorem 11 addresses the uniform case and Theorem 15 addresses the general case. Throughout this section, we let $L(\mathscr{X}_1, \ldots, \mathscr{X}_n)$ denote the length of the shortest tour that visits one element from each point set $\mathscr{X}_i$, which has $|\mathscr{X}_i| = k \geq 2$ for all i; see Figure 1 for an example.

Theorem 11 (Uniform demand)**.** *Let $k \geq 2$ be fixed and let $\mathscr{X}_1, \ldots, \mathscr{X}_n$ be point sets of cardinality k that are all drawn independently and uniformly at random in a region of unit area in $\mathbb{R}^2$. Then for all fixed $k \geq 2$, we have*

$$0.4839 =: \lambda_1 < \liminf_{n\to\infty} \frac{L(\mathscr{X}_1, \ldots, \mathscr{X}_n)}{\sqrt{n/k}} \leq \limsup_{n\to\infty} \frac{L(\mathscr{X}_1, \ldots, \mathscr{X}_n)}{\sqrt{n/k}} < \mu_1 := 1.8408 \qquad (5)$$

with probability one.

Proof. The upper bounding constant is conceptually very simple: first, take the special case where $\mathcal{R}$ is the unit square, and from each point set $\mathscr{X}_i$, let point X_i be the member of $\mathscr{X}_i$ that lies the farthest to the right. The points X_i follow the probability distribution

$$f(x = (x_1, x_2)) = k x_1^{k-1}$$

and Theorem 4 says that a TSP tour of a collection of points following this distribution must satisfy (with probability one)

$$\lim_{n\to\infty} \frac{\mathrm{TSP}(X_1, \ldots, X_n)}{\sqrt{n}} = \beta_2 \int_{\mathcal{R}} \sqrt{f(x)}\, \mathrm{d}x = \int_0^1 \int_0^1 \sqrt{k x_1^{k-1}}\, \mathrm{d}x_1\, \mathrm{d}x_2 = \frac{2\beta_2\sqrt{k}}{k+1} < 2\beta_2/\sqrt{k} \qquad (6)$$

where β_2 is the BHH constant, and so we merely apply the bound of $\beta_2 < 0.9204$ from Section 8.5 of [11] which gives the desired value of μ_1.

When $\mathcal{R}$ is an arbitrary region (of unit area), the intuition is the same, but we exploit the fact that the selection of the rightmost point in the previous construction was arbitrary (as opposed to, say, the leftmost point). Let $\mathcal{R}_1, \ldots, \mathcal{R}_s$ be a partition of $\mathcal{R}$ into measurable pieces with area $1/s$, and for each point set $\mathscr{X}_i$, let X_i be the member of $\mathscr{X}_i$ that belongs to the piece of the highest index (breaking ties arbitrarily). We have

$$\begin{aligned} &\mathsf{Pr}(X_i \in \bigcup_{j'=1}^{j} \mathcal{R}_{j'}) = \left(\frac{j}{s}\right)^k \\ \implies &\mathsf{Pr}(X_i \in \mathcal{R}_j) = \left(\frac{j}{s}\right)^k - \left(\frac{j-1}{s}\right)^k =: p_j \end{aligned}$$

with $\boldsymbol{p} \in \Delta^{s-1}$, the usual probability simplex. The X_i's follow a step distribution $\phi(x) = s\sum_{j=1}^{s} p_j \mathbb{1}\{x \in \mathcal{R}_j\}$, and so Theorem 4 says that a TSP tour of a collection of points following this distribution must satisfy (with probability one)

$$\begin{aligned} \lim_{n\to\infty} \frac{\mathsf{TSP}(X_1, \ldots, X_n)}{\sqrt{n}} = \beta_2 \int_{\mathcal{R}} \sqrt{\phi(x)}\, \mathsf{d}x &= \frac{\beta_2}{s} \sum_{j=1}^{s} \sqrt{s p_j} \\ &\to \beta_2 \int_0^1 \sqrt{k t^{k-1}}\, \mathsf{d}t \text{ as } s \to \infty \\ &= \frac{2\beta_2\sqrt{k}}{k+1} < 2\beta_2/\sqrt{k} \end{aligned}$$

as desired.

To derive the lower bound, let E_n be the event that $L(\mathscr{X}_1, \ldots, \mathscr{X}_n) < c\sqrt{n/k}$ for fixed c. Applying the union bound to Corollary 9 for the case $d = 2$ and using the fact that there are k^n different possible ways to select one member from each set $\mathscr{X}_i$, we see that

$$\begin{aligned} \mathsf{Pr}(E_n) &\leq k^n \cdot \frac{\Gamma(n+1)}{\Gamma(2n+1)} \left(\frac{2\pi c^2 n}{k}\right)^n \\ \implies \log \mathsf{Pr}(E_n) &\leq (1 + 2\log c - \log 2 + \log \pi)n - \mathcal{O}(1) \end{aligned} \tag{7}$$

where we have applied Lemma 3. We see that $(7) \to -\infty$ if and only if the coefficient of n is negative:

$$\begin{aligned} 0 &> 1 + 2\log c - \log 2 + \log \pi \\ &\Updownarrow \\ c &< \sqrt{\frac{2}{\pi e}} \approx 0.48393\,. \end{aligned} \tag{8}$$

Furthermore, this guarantees that $\mathsf{Pr}(E_n) \leq a^{-n}$ for some $a > 1$, so that $\sum_{n=1}^{\infty} \mathsf{Pr}(E_n) < \infty$. We apply Lemma 6 to obtain $\lambda_1 < \liminf_{n\to\infty} L(\mathscr{X}_1, \ldots, \mathscr{X}_n)/\sqrt{n/k}$ with probability one, which completes the proof. □

To attack the case where the samples arise from a non-uniform distribution, we require the following consequence of Theorem 11:

Corollary 12 (Tour length in a subset with uniform demand)**.** *Let $\mathscr{X}_1, \ldots, \mathscr{X}_n$ be independent uniform samples of cardinality k drawn from a compact region $\mathcal{R}$ with volume 1 and let $\mathcal{S} \subset \mathcal{R}$. Then*

$$\liminf_{n\to\infty} \frac{L(\mathscr{X}_i : \mathscr{X}_i \subset \mathcal{S})}{\sqrt{n}} > \lambda_1 \sqrt{\frac{\text{Area}(\mathcal{S})^{k+1}}{k}}$$

with probability one, where $\mathscr{X}_i \subset \mathcal{S}$ means that all k elements of $\mathscr{X}_i$ lie in $\mathcal{S}$ and $\lambda_1 = 0.4839$ from Theorem 11.

Proof. Let $\{\mathscr{Y}_i\}$ denote a sequence of uniform samples of cardinality k drawn from $\mathcal{S}$ (not $\mathcal{R}$). Certainly, by scaling areas, Theorem 11 says that

$$\liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n)}{\sqrt{n}} > \lambda_1 \sqrt{\text{Area}(\mathcal{S})/k}\,.$$

We now define $N(n) = |i \in \{1, \ldots, n\} : \mathscr{X}_i \subset \mathcal{S}|$ and $p = \text{Area}(\mathcal{S})^k$ to be the probability that $\mathscr{X}_i \subset \mathcal{S}$. We have (with probability 1)

$$\begin{aligned}
\liminf_{n\to\infty} \frac{L(\mathscr{X}_i : \mathscr{X}_i \subset \mathcal{S})}{\sqrt{n}} &= \liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_{N(n)})}{\sqrt{n}} \\
&= \liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_{N(n)})}{\sqrt{n}} \cdot \sqrt{\frac{N(n)}{N(n)}} \\
&= \liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_{N(n)})}{\sqrt{N(n)}} \cdot \sqrt{\frac{N(n)}{n}} \\
&= \liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_{N(n)})}{\sqrt{N(n)}} \left(\lim_{n\to\infty} \sqrt{\frac{N(n)}{n}} \right) \left(\right. \\
&> \left(\lambda_1 \sqrt{\text{Area}(\mathcal{S})/k} \right) \sqrt{p} = \lambda_1 \sqrt{\text{Area}(\mathcal{S})^{k+1}/k}
\end{aligned}$$

as desired. □

We next prove the non-uniform convergence result for the special case where the density is a step function:

Lemma 13 (Tour length from a step density)**.** *Let $\phi(x) = \sum_{i=1}^{s} a_i \mathbb{1}(x \in \boxdot_i)$ be a step density function with compact support $\mathcal{R}$ such that $a_1 \geq \cdots \geq a_s$ and a_i Area$(\boxdot_i) = 1/s$ for all i (so that* Area$(\mathcal{R}) = 1$*). If $\mathscr{Y}_1, \ldots, \mathscr{Y}_n$ are independent samples from ϕ having cardinality k and $\Pi(x)$ is defined as in Lemma 10, then*

$$\liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n)}{\sqrt{n}} \geq \frac{\lambda_1 (k+1)}{2\sqrt{k}} \int_{\mathcal{R}} \sqrt{\phi(x)\Pi(x)^{k-1}}\, \mathrm{d}x$$

with probability one, where $\lambda_1 = 0.4839$ from Theorem 11.

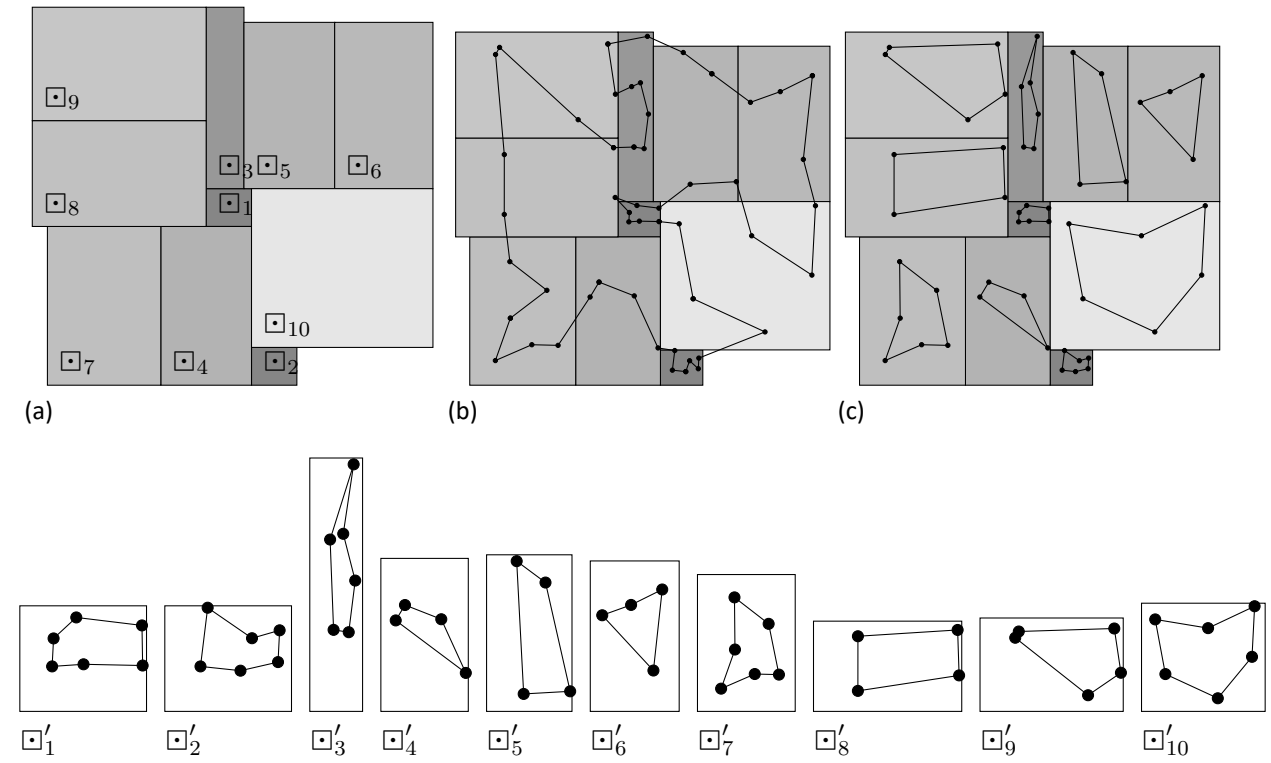


Figure 2: Figure 2a shows a step function ϕ that presumably satisfies the conditions of Lemma 13 (in the sense that darker – i.e. denser – regions are smaller, reflecting the assumption that a_i Area$(\boxdot_i) = 1/s$ for all i). Figure 2b shows the TSP tour of a collection of independent samples $Y_1, \ldots, Y_n$ of ϕ, whose length differs from that of a collection of tours within each component (Figure 2c) by a constant, by Lemma 5. Figure 2d shows the re-scaled components $\boxdot'_i = \Psi(\boxdot_i)$ and the points $Y'_1, \ldots, Y'_n$.

Proof. We have

$$\begin{aligned} L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n) &= \min_{Y_i \in \mathscr{Y}_i} \mathsf{TSP}(Y_1, \ldots, Y_n) \\ &= \min_{Y_i \in \mathscr{Y}_i} \sum_{i=1}^{s} \Big(\mathsf{TSP}(Y_1, \ldots, Y_n \cap \boxdot_i) + \mathcal{O}(1) \end{aligned}$$

from Lemma 5. For each i, define $\Psi_i : \boxdot_i \to \mathbb{R}^2$ by $\Psi_i(y) = \sqrt{a_i} y + \xi_i$, where the ξ_i are selected so that the images $\Psi_i(\boxdot_i)$ are all disjoint (their specific values are irrelevant). Let $\Psi : \mathcal{R} \to \mathbb{R}^2$ be the union of all the Ψ_i's (i.e. for any $y \in \mathcal{R}$, we have $\Psi(y) = \Psi_i(y)$, where $\boxdot_i \ni y$). The significance of this construction is that the image $\Psi(\mathcal{R})$ has area 1 and $\Psi(\mathscr{Y}_1), \ldots, \Psi(\mathscr{Y}_n)$ becomes a uniform collection of samples in $\Psi(\mathcal{R})$. For notational compactness, define $Y'_i = \Psi(Y_i)$, $\mathscr{Y}'_i = \Psi(\mathscr{Y}_i)$, and $\boxdot'_i = \Psi(\boxdot_i) = \Psi_i(\boxdot_i)$; Figure 2 shows this construction. Basic scaling arguments tell us that

$$\mathsf{TSP}(\{Y_1, \ldots, Y_n\} \cap \boxdot_i) = \frac{1}{\sqrt{a_i}} \mathsf{TSP}(\{Y'_1, \ldots, Y'_n\} \cap \boxdot'_i) \,.$$

Since the a_i's are decreasing, we can define increasing terms $b_i = 1/\sqrt{a_i}$ and we can also construct non-negative c_j's so that $b_i = \sum_{j=1}^{i} c_j$. This tells us that

$$\begin{aligned}
\min_{Y_i \in \mathscr{Y}_i} \sum_{i=1}^{s} \mathsf{TSP}(\{Y_1, \ldots, Y_n\} \cap \boxdot_i) &= \min_{Y_i' \in \mathscr{Y}_i'} \sum_{i=1}^{s} \left(\frac{1}{\sqrt{a_i}} \mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_i') \right) \\
&= \min_{Y_i' \in \mathscr{Y}_i'} \sum_{i=1}^{s} \left(b_i \mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_i') \right) \\
&= \min_{Y_i' \in \mathscr{Y}_i'} \sum_{i=1}^{s} \left(\sum_{j=1}^{i} c_j \right) \left(\mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_i') \right) \\
&= \min_{Y_i' \in \mathscr{Y}_i'} \sum_{j=1}^{s} c_j \sum_{i=j}^{s} \left(\mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_i') \right)
\end{aligned}$$

and now note that for all j, Lemma 5 also tells us that

$$\sum_{i=j}^{s} \left(\mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_i') = \mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_{\geq j}') + \mathcal{O}(1) \right),$$

where we define $\boxdot_{\geq j}' = \bigcup_{i=j}^{s} \left(\boxdot_i' \right)$. It is certainly true that

$$\mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \boxdot_{\geq j}') \geq L(\mathscr{Y}_p' : \mathscr{Y}_p' \subset \boxdot_{\geq j}')$$

because if $\mathscr{Y}_p' \subset \boxdot_{\geq j}'$, then certainly $Y_p' \in \boxdot_{\geq j}'$. The samples $\mathscr{Y}_p' : \mathscr{Y}_p' \subset \boxdot_{\geq j}'$ are independently and uniformly distributed within $\boxdot_{\geq j}'$. Therefore, since $\text{Area}(\boxdot_{\geq j}') = (s - j + 1)/s$, Lemma 12 says that

$$\liminf_{n \to \infty} \frac{1}{\sqrt{n}} L(\mathscr{Y}_p' : \mathscr{Y}_p' \subset \boxdot_{\geq j}') \geq \lambda_1 \sqrt{\left(\frac{s - j + 1}{s} \right)^{k+1} / k}$$

and so if we adopt the convention that $b_0 = 0$, we have

$$\begin{aligned}
\liminf_{n \to \infty} \frac{1}{\sqrt{n}} \min_{Y_i \in \mathscr{Y}_i} \sum_{j=1}^{s} c_j \sum_{i=j}^{s} \left(\mathsf{TSP}(\{Y_1', \ldots, Y_n'\} \cap \sqrt{a_i} \boxdot_i) \right) &\geq \frac{\lambda_1}{\sqrt{k}} \sum_{j=1}^{s} c_j \sqrt{\left(\frac{s - j + 1}{s} \right)^{k+1}} \\
&= \frac{\lambda_1}{\sqrt{k}} \sum_{j=1}^{s} (b_j - b_{j-1}) \sqrt{\left(\frac{s - j + 1}{s} \right)^{k+1}} \\
&\geq \frac{\lambda_1}{\sqrt{k s^{k+1}}} \sum_{j=1}^{s} b_j \left[\sqrt{(s - j + 1)^{k+1}} - \sqrt{(s - j)^{k+1}} \right]
\end{aligned}$$

and we have

$$\sqrt{(s - j + 1)^{k+1}} - \sqrt{(s - j)^{k+1}} \geq \frac{k+1}{2} \sqrt{(s - j)^{k-1}}$$

for all $j \le s$ and k, thus

$$\begin{aligned}
&\frac{\lambda_1}{\sqrt{k s^{k+1}}} \sum_{j=1}^{s} b_j \left[\sqrt{(s-j+1)^{k+1}} - \sqrt{(s-j)^{k+1}} \right] \\
&\ge \frac{\lambda_1 (k+1)}{2\sqrt{k s^{k+1}}} \sum_{j=1}^{s} b_j \sqrt{(s-j)^{k-1}} \\
&= \frac{\lambda_1 (k+1)}{2s\sqrt{k}} \sum_{j=1}^{s} \frac{1}{\sqrt{a_j}} \sqrt{\left(\frac{s-j}{s}\right)^{k-1}} \\
&= \frac{\lambda_1 (k+1)}{2\sqrt{k}} \sum_{j=1}^{s} \sqrt{a_j}\, \mathsf{Area}(\boxdot_j) \sqrt{\left(\frac{s-j}{s}\right)^{k-1}} \\
&= \frac{\lambda_1 (k+1)}{2\sqrt{k}} \sum_{j=2}^{s+1} \sqrt{a_j}\, \mathsf{Area}(\boxdot_j) \sqrt{\left(\frac{s-j+1}{s}\right)^{k-1}} \\
&= \frac{\lambda_1 (k+1)}{2\sqrt{k}} \sum_{j=2}^{s} \sqrt{a_j}\, \mathsf{Area}(\boxdot_j) \sqrt{\left(\frac{s-j+1}{s}\right)^{k-1}} \\
&= \frac{\lambda_1 (k+1)}{2\sqrt{k}} \left(\sum_{j=1}^{s} \sqrt{a_j}\, \mathsf{Area}(\boxdot_j) \sqrt{\left(\frac{s-j+1}{s}\right)^{k-1}} - \sqrt{a_1}\, \mathsf{Area}(\boxdot_1) \right) \\
&\ge \frac{\lambda_1 (k+1)}{2\sqrt{k}} \left(\sum_{j=1}^{s} \sqrt{a_j}\, \mathsf{Area}(\boxdot_j) \sqrt{\left(\frac{s-j+1}{s}\right)^{k-1}} - \frac{1}{s}\sqrt{\mathsf{Area}(\mathcal{R})} \right) \qquad (9) \\
&= \frac{\lambda_1 (k+1)}{2\sqrt{k}} \int_{\mathcal{R}} \sqrt{\phi(x)\Pi(x)^{k-1}}\, \mathrm{d}x - \underbrace{\frac{\lambda_1 (k+1)}{2s\sqrt{k}} \sqrt{\mathsf{Area}(\mathcal{R})}}_{(*)} \qquad (10)
\end{aligned}$$

as desired, where (9) uses the fact that $a_1\, \mathsf{Area}(\boxdot_1) = 1/s$ and a_1 is the largest of all the a_i's. The desired result follows from the fact that we can make s as large as we like by breaking each component $\boxdot_i$ of ϕ into multiple components of equal area without changing the function ϕ, which allows us to drop the term $(*)$. □

To attack the non-uniform case of Theorem 11, we require one more observation:

Claim 14. If $\int_{\mathcal{R}} |g|\, \mathrm{d}x \le \delta$, then $\int_{\mathcal{R}} \sqrt{|g|}\, \mathrm{d}x \le \sqrt{\mathsf{Area}(\mathcal{R})\delta}$.

Proof. If we maximize $\int_{\mathcal{R}} \sqrt{g}\, \mathrm{d}x$ subject to the constraint that $\int_{\mathcal{R}} g\, \mathrm{d}x \le \delta$, the solution g^* is uniform with $g^*(x) = \delta / \mathsf{Area}(\mathcal{R})$ everywhere. □

It is now a simple matter to apply Lemma 10:

Theorem 15. *Let f, $\mathcal{R}$, and P be as in Lemma 10. With probability one, we have*

$$\frac{\lambda_1}{2} < \liminf_{n\to\infty} \frac{L(\mathscr{X}_1, \ldots, \mathscr{X}_n)}{\sqrt{nk} \int_{\mathcal{R}} \sqrt{f(x)P(x)^{k-1}}\, \mathrm{d}x} \le \limsup_{n\to\infty} \frac{L(\mathscr{X}_1, \ldots, \mathscr{X}_n)}{\sqrt{nk} \int_{\mathcal{R}} \sqrt{f(x)P(x)^{k-1}}\, \mathrm{d}x} < \frac{\mu_1}{2} \qquad (11)$$

with λ_1, μ_1 as in Theorem 11.

Proof. The upper bounding argument is very simple: for each point set $\mathscr{X}_i$, let point X_i be the member of $\mathscr{X}_i$ where f is the densest, i.e. $X_i = \arg\max_{x \in \mathscr{X}_i} f(x)$. This is intuitive because it is desirable to select those points that belong to more densely populated areas. The density function on the X_i's is

$$f(x) = kf(x)P(x)^{k-1}$$

and by Theorem 4, the length of a tour of those X_i's satisfies

$$\lim_{n\to\infty} \frac{\mathrm{TSP}(X_1, \ldots, X_n)}{\sqrt{n}} = \beta_2 \int_{\mathcal{R}} \sqrt{f(x)}\, \mathrm{d}x = \beta_2 \iint_{\mathcal{R}} \sqrt{kf(x)P(x)^{k-1}}\, \mathrm{d}x \leq \frac{\mu_1}{2} \iint_{\mathcal{R}} \sqrt{kf(x)P(x)^{k-1}}\, \mathrm{d}x$$

where β_2 is the BHH constant.

To prove the lower bound, let ϕ be the approximation of f from Lemma 10. By a standard coupling argument [25], there is a joint distribution for random variables (X, Y) such that X has density f, Y has density ϕ, and $\Pr(X \neq Y) \leq \epsilon/k$ for any ϵ; this means that if $\mathscr{X} = (X_1, \ldots, X_k)$ is a collection of k independent samples of f and $\mathscr{Y} = (Y_1, \ldots, Y_k)$ is a collection of k independent samples of ϕ, then $\Pr(\mathscr{X} \neq \mathscr{Y}) < \epsilon$. We have

$$\begin{aligned}
L(\mathscr{X}_1, \ldots, \mathscr{X}_n) &\geq L(\mathscr{X}_1, \ldots, \mathscr{X}_n : \mathscr{X}_i = \mathscr{Y}_i) \\
&= L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n : \mathscr{X}_i = \mathscr{Y}_i) \\
&\geq L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n) - L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n : \mathscr{X}_i \neq \mathscr{Y}_i) - \mathcal{O}(1) \\
\implies \liminf_{n\to\infty} \frac{L(\mathscr{X}_1, \ldots, \mathscr{X}_n)}{\sqrt{n}} &\geq \liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n)}{\sqrt{n}} - \frac{\alpha_2 \sqrt{\mathrm{Area}(\mathcal{R})\epsilon n}}{\sqrt{n}} \\
&= \liminf_{n\to\infty} \frac{L(\mathscr{Y}_1, \ldots, \mathscr{Y}_n)}{\sqrt{n}} - \alpha_2 \sqrt{\epsilon\, \mathrm{Area}(\mathcal{R})} \\
&\geq \frac{\lambda_1(k+1)}{2\sqrt{k}} \iint_{\mathcal{R}} \sqrt{\phi(x)\Pi(x)^{k-1}}\, \mathrm{d}x - \alpha_2 \sqrt{\epsilon\, \mathrm{Area}(\mathcal{R})}
\end{aligned}$$

where we have applied Lemma 13 in the last inequality. Finally, note that we can select our approximation ϕ, Π arbitrarily closely so that, by Lemma 10 (compactifying our notation momentarily),

$$\begin{aligned}
\epsilon &\geq \max_x \; P(x)^{k-1} - \Pi(x)^{k-1} \; + \iint_{\mathcal{R}} |\phi(x) - f(x)|\, \mathrm{d}x \\
&\geq \int \phi \; P^{k-1} - \Pi^{k-1} \; + \iint |\phi - f|\, P^{k-1} \\
&\geq \iint \phi\Pi^{k-1} - \phi P^{k-1} + \phi P^{k-1} - f P^{k-1} \\
&= \iint \phi\Pi^{k-1} - f P^{k-1} \\
\implies \sqrt{\mathrm{Area}(\mathcal{R})\epsilon} &\geq \iint \sqrt{|\phi\Pi^{k-1} - f P^{k-1}|} \geq \int \sqrt{\phi\Pi^{k-1}} - \iint \sqrt{f P^{k-1}} \\
\implies \int \sqrt{\phi\Pi^{k-1}} &\geq \iint \sqrt{f P^{k-1}} - \sqrt{\mathrm{Area}(\mathcal{R})\epsilon}
\end{aligned}$$

and therefore, we ultimately conclude that

$$\begin{aligned}\liminf_{n\to\infty} \frac{L(\mathcal{X}_1,\ldots,\mathcal{X}_n)}{\sqrt{n}} &\geq \frac{\lambda_1(k+1)}{2\sqrt{k}} \int_{\mathcal{R}} \sqrt{\phi(x)\Pi(x)^{k-1}}\,\mathrm{d}x - \alpha_2\sqrt{\epsilon\,\mathsf{Area}(\mathcal{R})} \\ &\geq \frac{\lambda_1(k+1)}{2\sqrt{k}} \left(\int_{\mathcal{R}} \sqrt{f(x)P(x)^{k-1}}\,\mathrm{d}x - \sqrt{\mathsf{Area}(\mathcal{R})\epsilon}\right) - \alpha_2\sqrt{\epsilon\,\mathsf{Area}(\mathcal{R})} \\ &\geq \frac{\lambda_1\sqrt{k}}{2} \int_{\mathcal{R}} \sqrt{f(x)P(x)^{k-1}}\,\mathrm{d}x - \underbrace{\sqrt{\epsilon\,\mathsf{Area}(\mathcal{R})}\left(\frac{\lambda_1(k+1)}{2\sqrt{k}} - \alpha_2\right)}_{(*)} \qquad (12)\end{aligned}$$

which completes the lower bound (and therefore, the proof), since (12) can be made as small as desired by choosing small values of ϵ. □

6.1 The GTSP with clustering

In this section we study the case where samples are not independently drawn; rather, we will assume that each point set $\mathcal{X}_i$ consists of points that are clustered together. This assumption holds in the vast majority of instances of the GTSP; for example, the benchmark data in [12] is constructed by selecting problems from the TSPLIB library [29] and then grouping point sets together based on proximity. Our motivation for doing so is in order to derive some managerial insights before proceeding to the next selection routing problem of interest.

Our model of clustering is as follows: we assume that we are given a fixed (compact) Jordan measurable shape $\mathcal{S}$ of arbitrary volume, and that each $\mathcal{X}_i$ is obtained by placing $\mathcal{S}$ uniformly at random in the unit cube $[0,1]^d$ and sampling k points uniformly within $\mathcal{S}$. In order to sidestep boundary effects that might occur by having $\mathcal{S}$ only partially contained in the cube, we assume that the uniform placement of $\mathcal{S}$ is done in a "toroidal" fashion (in which opposing sides of the cube are "glued" together), as suggested in Figure 3. It turns out that the bounds from Theorem 11 remain valid even in this situation:

Theorem 16. *The bounds from Theorem 11 remain valid when point sets $\mathcal{X}_i$ are sampled from a uniformly placed shape $\mathcal{S}$ as described above.*

Proof. We first note that, if one selects points $X_1 \in \mathcal{X}_1, \ldots, X_n \in \mathcal{X}_n$ arbitrarily, then the resulting samples $X_1,\ldots,X_n$ are still uniformly and independently drawn in the cube, by virtue of the fact that $\mathcal{S}$ was placed uniformly at random. Hence, we can again apply the union bound to Corollary 9 exactly as in the proof of Theorem 11, so that our lower bound is unaffected. Thus, our proof is complete if we can show that the upper bounds b_d apply as well. Note that in Theorem 11, we derived those upper bounds by selecting the member of each $\mathcal{X}_i$ whose first entry was the smallest. It is obvious that this is not guaranteed to work here; consider the case where $\mathcal{S}$ is a ball whose radius is very small.

Since $\mathcal{S}$ is Jordan measurable, it can be approximated to arbitrary precision by a finite set of disjoint rectangles. More precisely, for any threshold $\epsilon' > 0$, we can construct a finite set of disjoint rectangles $R_1,\ldots,R_m$ (where m varies depending on ϵ'), each of which is contained in $\mathcal{S}$, such that

$$\sum_{i=1}^{m} \mathsf{Vol}(R_i) \geq \mathsf{Vol}(\mathcal{S}) - \epsilon'\,.$$

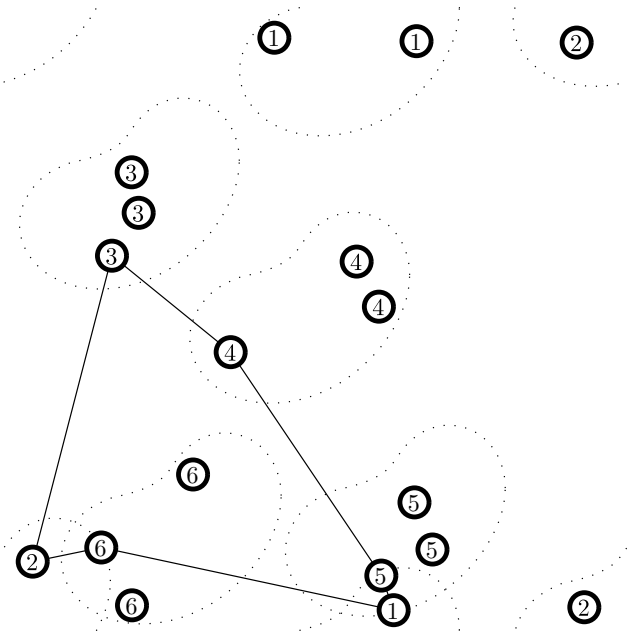


Figure 3: A generalized TSP tour of six sets of points $\mathcal{X}_1, \ldots, \mathcal{X}_6$, each consisting of $k = 3$ points, where $\mathcal{S}$ is a "jellybean" type shape that is placed uniformly at random in a "toroidal" fashion in the unit square.

We now fix a threshold $\epsilon > 0$, and let $R_1, \ldots, R_m$ be disjoint rectangles contained in $\mathcal{S}$ such that

$$\sum_{i=1}^{m} \text{Vol}(R_i) \geq \left[1 - \left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]^{1/(2k)} \text{Vol}(\mathcal{S}),$$

and we define $\mathcal{R} = \bigcup_{i=1}^{m} R_i$. Next, we write the dimensions of each rectangle R_i as $s_1^i \times s_2^i \times \cdots \times s_d^i$, we define $s = \min_i\{s_1^i\}$, and we define $\delta > 0$ to be any number such that $1/\delta$ is an integer and

$$\delta \leq \left(1 - \left[1 - \left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]^{1/(2k)}\right) \frac{s}{2}.$$

Our "algorithm" for selecting one of the k samples drawn from $\mathcal{S}$ is described below:

1. Place $\mathcal{S}$ uniformly at random (in a "toroidal" fashion) in the unit cube $[0,1]^d$, and place $\mathcal{R}$ correspondingly in $\mathcal{S}$. Draw k samples $X_1, \ldots, X_k$ uniformly at random from $\mathcal{S}$.

2. Let each $R_i \in \mathcal{R}$ be written as

   $$R_i = [u_1^i, v_1^i] \times [u_2^i, v_2^i] \times \cdots \times [u_d^i, v_d^i]$$

   and define $\tilde{R}_1, \ldots, \tilde{R}_m$ by

   $$\tilde{R}_i = \left[\left\lceil \frac{u_1^i}{\delta} \right\rceil \delta, \left\lfloor \frac{v_1^i}{\delta} \right\rfloor \delta\right] \times [u_2^i, v_2^i] \times \cdots \times [u_d^i, v_d^i];$$

   that is, round u_1^i upwards and round v_1^i downwards to their nearest integer multiple of δ (this obviously implies that $\tilde{R}_i \subseteq R_i$). Define $\tilde{\mathcal{R}} = \bigcup_{i=1}^{m} \tilde{R}_i$.

3. If any one of the k samples $X_1, \ldots, X_k$ lies outside $\tilde{\mathcal{R}}$, set $X^* = X_1$ (i.e. an arbitrary choice) and return X^*. Otherwise, go to step 4.

4. Write each sample X_i as $X_i = (x_1^i, \ldots, x_d^i)$. For each sample X_i, define

$$t_i = x_1^i - \left\lfloor \frac{x_1^i}{\delta} \right\rfloor \delta$$

and let X^* be the sample for which t_i is minimal. Return X^*.

The salient properties of the above scheme are as follows: on the one hand, if any of the k samples lies outside $\tilde{\mathcal{R}}$, then it is obvious that the returned sample X^* simply follows a uniform distribution on $[0,1]^d$ (since $\mathcal{S}$ was positioned uniformly at random). On the other hand, if all k samples lie in $\tilde{\mathcal{R}}$, then it is not hard to see that each value t_i is uniformly distributed between 0 and δ, with all t_i's being independent; this is because the rounding that defines the $\tilde{R}_i$'s in step 2 guarantees that each x_1^i is a uniform sample from an interval whose endpoints are integer multiples of δ. Therefore, the minimum of the t_i's follows a probability density function $h(t)$ given by

$$h(t) = \frac{k(\delta - t)^{k-1}}{\delta^k},$$

and so the selection X^* follows a probability density $g(\mathbf{x})$ given by

$$g(\mathbf{x}) = \frac{k[\delta - (x_1 - \lfloor x_1/\delta \rfloor\, \delta)]^{k-1}}{\delta^{k-1}},$$

as shown in Figure 4. This density function is periodic (with period δ) with respect to the first coordinate x_1 of $\mathbf{x}$ and is concentrated near one side of the hyperplanes of the form $x_1 = q\delta$ with $q \in \mathbb{Z}$. Note that if one draws n' samples $X_1^*, \ldots, X_{n'}^*$ from this distribution, the BHH theorem says that their TSP tour must be almost surely asymptotic to

$$\beta_d n'^{(d-1)/d} \int\!\!\cdots\!\!\int_{[0,1]^d} g(\mathbf{x})^{(d-1)/d}\, d\mathbf{x} = \beta_d n'^{(d-1)/d} \frac{1}{\delta} \int_0^\delta \left[\frac{k(\delta - t)^{k-1}}{\delta^{k-1}}\right]^{(d-1)/d} dt = \beta_d n'^{(d-1)/d} \cdot \frac{dk^{(d-1)/d}}{(d-1)k+1}$$
$$< \frac{\beta_d d}{d-1} \left(\frac{n'^{d-1}}{k}\right)^{1/d}$$

which is the same expression as (6).

To put all of the pieces together, we let $f(\mathbf{x})$ denote the probability density function associated with the sample X^* obtained according to our "algorithm". It is easy to see that $f(\mathbf{x})$ is a mixture of the density $g(\mathbf{x})$ that we just defined together with a uniform distribution, which we will write as $f(\mathbf{x}) = \lambda g(\mathbf{x}) + (1 - \lambda)$ for some $\lambda \in [0,1]$. We want to show that λ is close to 1. In step 2, note that each rectangle $\tilde{R}_i$ is obtained by increasing u_1^i by less than δ and decreasing v_1^i by less than δ, and consequently

$$\text{Vol}(\tilde{R}_i) > \frac{v_1^i - u_1^i - 2\delta}{v_1^i - u_1^i} \text{Vol}(R_i) \geq \frac{s - 2\delta}{s} \text{Vol}(R_i),$$

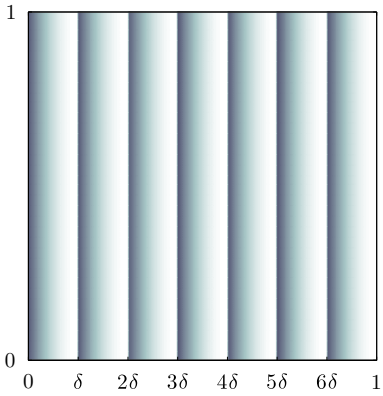


Figure 4: The function $g(\mathbf{x})$ in $[0,1]^2$, with $\delta = 1/7$. We have used $k = 3$, although this is not reflected in any obvious way by the shading above.

where $s = \min_i\{s_1^i\}$ as defined above, and therefore

$$\begin{aligned} \text{Vol}(\tilde{\mathcal{R}}) = \sum_{i=1}^{m} \text{Vol}(\tilde{R}_i) > \frac{s-2\delta}{s} \sum_{i=1}^{m} \text{Vol}(R_i) &\geq \frac{s-2\delta}{s}\left[1-\left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]\text{Vol}(\mathcal{S}) \\ \implies \frac{\text{Vol}(\tilde{\mathcal{R}})}{\text{Vol}(\mathcal{S})} \geq \frac{s-2\delta}{s}\left[1-\left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]^{1/(2k)} &\geq \left[1-\left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]^{1/k}. \end{aligned}$$

Therefore, for any placement of $\mathcal{S}$, the probability that all k samples $X_1, \ldots, X_k$ lie in $\tilde{\mathcal{R}}$ is at least

$$\left[1-\left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]^{1/k}\Bigg)^k = 1-\left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)},$$

which says that $\lambda \geq 1-(\epsilon/k^{1/d})^{d/(d-1)}$. Thus, by the BHH theorem, the TSP tour of a collection of n points sampled from the distribution $f(\mathbf{x})$ is proportional to

$$\begin{aligned} \beta_d n^{(d-1)/d}\int_{[0,1]^d} f(\mathbf{x})^{(d-1)/d}\,d\mathbf{x} &= \beta_d n^{(d-1)/d}\int_{[0,1]^d}[\lambda g(\mathbf{x})+(1-\lambda)]^{(d-1)/d}\,d\mathbf{x} \\ &\leq \beta_d n^{(d-1)/d}\int_{[0,1]^d}[g(\mathbf{x})+(1-\lambda)]^{(d-1)/d}\,d\mathbf{x} \\ &\leq \beta_d n^{(d-1)/d}\int_{[0,1]^d}\left[g(\mathbf{x})+\left(\frac{\epsilon}{k^{1/d}}\right)^{d/(d-1)}\right]^{(d-1)/d}\,d\mathbf{x} \\ &\leq \beta_d n^{(d-1)/d}\int_{[0,1]^d} g(\mathbf{x})^{(d-1)/d}+\frac{\epsilon}{k^{1/d}}\,d\mathbf{x} \\ &= \frac{\beta_d d}{d-1}\left(\frac{n^{d-1}}{k}\right)^{1/d}+\epsilon\frac{\beta_d n^{(d-1)/d}}{k^{1/d}} \\ &< \frac{\beta_d d}{d-1}\left(\frac{n^{d-1}}{k}\right)^{1/d}+\epsilon\frac{\beta_d d}{d-1}\left(\frac{n^{d-1}}{k}\right)^{1/d} \\ &= (1+\epsilon)\frac{\beta_d d}{d-1}\left(\frac{n^{d-1}}{k}\right)^{1/d}. \end{aligned}$$

Since ϵ was chosen arbitrarily, this completes the proof. □

6.2 Managerial insights for the GTSP

The preceding result is somewhat counterintuitive because it effectively says that the GTSP is unaffected by clustering within subsets, provided that the overall placement of the subsets themselves remains uniform and that the number of subsets is large. Managerially speaking, to refer back to the introduction, this implies that:

- If one employs a fully randomized distribution strategy in a warehouse, in which each SKU is uniformly randomly positioned at k separate locations, then the average length of an order-picking tour of n distinct SKUs is proportional to $\sqrt{n/k}$, as opposed to $\sqrt{n}$ in the deterministic case as established in [18]. Hence, one expects a reduction in total distance travelled that is proportional to $1/\sqrt{k}$.
- If one employs a randomized strategy in a warehouse, such as Amazon's random stow [8], Theorem 16 says that one does not need to distribute identical SKUs throughout the entire warehouse to reap the benefits of the GTSP; it suffices to randomize the SKU placement by selecting a small portion of the warehouse uniformly at random, then distributing the identical SKUs within it.
- The paper [6] uses the GTSP to study whether delivery services are guaranteed to improve the carbon footprint in a region; they find that, for households that already drive to many different locations that are all independently and uniformly distributed, the carbon footprint may actually increase. Theorem 16 says that this remains true when those locations are clustered near each other, which is a more realistic assumption (since Hotelling-style competition usually implies that similar businesses will locate close to one another).
- If one is using a socially networked system such as that described in [27, 38] (in which parcels can be delivered to multiple locations such as a customer's workplace or gym, in addition to their home), Theorem 16 says that one will see increases in system efficiency even when a customer's locations are all grouped near each other.

Remark 17. We emphasize that the assumption of "toroidal" placement is done for the sole purpose of ensuring that an arbitrary selection $X_i \in \mathcal{X}_i$ is still uniform, so that the lower bound from Theorem 11 is in effect; if one makes some other assumption about the behavior near the boundary, one might see small non-uniformities around the boundary of the unit cube. Clearly, when $\mathcal{S}$ is small, the likelihood of being placed near the boundary of the cube becomes small anyway, so the assumption is only an unrealistic one when $\mathcal{S}$ is large (in which case the amount of clustering within sets is expected to be small anyway).

7 Continuous approximation analysis of the cardinality constrained TSP (CCTSP)

This section gives a continuous approximation formula for the *cardinality constrainted TSP* (CCTSP). The CCTSP is an optimization problem in which we are given a set of points $X_1, \ldots, X_n$ and an inte-

ger $m < n$, and the goal is to find the shortest route that visits any m of the n points. Throughout this section, we let $L(X_1, \ldots, X_n; m)$ denote the length of the shortest tour that visits m of the points $X_1, \ldots, X_n$. Theorem 18 addresses the uniform case and Theorem 21 addresses the general case.

To save wear and tear on floors and ceilings, when m is non-integer, we round it upwards.

Theorem 18. *Let $X_1, \ldots, X_n$ be independent uniform samples drawn from a region of unit area in $\mathbb{R}^2$. For all fixed $0 < p < 1$, we have*

$$0.2935 =: \lambda_2 < \liminf_{n\to\infty} \frac{L(X_1, \ldots, X_n; pn)}{p\sqrt{n}} \leq \limsup_{n\to\infty} \frac{L(X_1, \ldots, X_n; pn)}{p\sqrt{n}} < \mu_2 := 0.9204 \quad (13)$$

with probability one.

Proof. The upper bounding constant is simple: take a TSP tour T of all of the points, whose length satisfies $\mathrm{TSP}(X_1, \ldots, X_n)/\sqrt{n} \to \beta_2$ with probability one (where β_2 is the BHH constant from Theorem 4). Fix an orientation of the tour and let T_i denote the subtour of T that begins at point i and traverses T until it has visited $\lceil pn \rceil$ points. Certainly,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathsf{length}(T_i) = (\lceil pn \rceil - 1)\, \mathsf{length}(T) \\
\implies &\min_i \mathsf{length}(T_i) \leq \frac{1}{n} \sum_{i=1}^{n} \mathsf{length}(T_i) = \frac{(\lceil pn \rceil - 1)}{n}\, \mathsf{length}(T) \\
\implies &\limsup_{n\to\infty} \frac{\min_i \mathsf{length}(T_i)}{p\sqrt{n}} \leq \lim_{n\to\infty} \frac{(\lceil pn \rceil - 1)}{n} \cdot \frac{\mathsf{length}(T)}{\sqrt{n}} = \beta_2 p < \mu_2 p
\end{aligned}$$

as desired, where we apply the bound of $\beta_2 < 0.9204$ from Section 8.5 of [11].

To derive the lower bound, let E_n be the event that $L(X_1, \ldots, X_n; pn) < cp\sqrt{n}$ for fixed c. We will apply the union bound to Corollary 9 for the case $d = 2$. Note that the number of possible subsets of cardinality $\lceil pn \rceil$ is $\binom{n}{\lceil pn \rceil}$, which satisfies

$$\begin{aligned}
\binom{n}{\lceil pn \rceil} &= \frac{\Gamma(n+1)}{\Gamma(\lceil pn \rceil + 1)\Gamma(n - \lceil pn \rceil + 1)} \\
&\leq \frac{\Gamma(n+1)}{\Gamma(pn+1)\Gamma(qn+2)} = \frac{1}{qn+1} \cdot \frac{\Gamma(n+1)}{\Gamma(pn+1)\Gamma(qn+1)} \\
\implies \log \binom{n}{\lceil pn \rceil} &\leq \log\Gamma(n+1) - \log\Gamma(pn+1) - \log\Gamma(qn+1) - \log(qn+1) \\
&= -(p\log p + q\log q)\, n + \mathcal{O}(\log n)
\end{aligned}$$

and so

$$\begin{aligned}
\Pr(E_n) &\leq \binom{n}{\lceil pn \rceil} \frac{\Gamma(\lceil pn \rceil + 1)}{\Gamma(2\lceil pn \rceil + 1)} \left(2\pi c^2 p^2 n\right)^{pn} \\
&\leq \binom{n}{\lceil pn \rceil} \frac{\Gamma(pn+2)}{\Gamma(2pn+1)} \left(2\pi c^2 p^2 n\right)^{pn} = \binom{n}{\lceil pn \rceil} (pn+1) \frac{\Gamma(pn+1)}{\Gamma(2pn+1)} \left(2\pi c^2 p^2 n\right)^{pn} \\
\implies \log \Pr(E_n) &\leq (2p\log c + p - q\log q - p\log 2 + p\log\pi)\, n + \mathcal{O}(\log n)\,.
\end{aligned}$$

The above expression approaches $-\infty$ if and only if the coefficient of n is negative:

$$\begin{aligned} 0 &> 2p\log c + p - q\log q - p\log 2 + p\log\pi \\ &\Updownarrow \\ c &< \sqrt{\frac{2}{\pi e}\cdot q^{q/p}} = \sqrt{\frac{2}{\pi e}\cdot (1-p)^{(1-p)/p}}\,. \end{aligned}$$

this is convex and increasing in p, and satisfies

$$\lim_{p\to 0^+}\sqrt{\frac{2}{\pi e}\cdot (1-p)^{(1-p)/p}} = \frac{\sqrt{2\pi}}{e} > 0.2935 =: \lambda_2\,.$$

Furthermore, this guarantees that $\Pr(E_n) \le a^{-n}$ for some $a > 1$, so that $\sum_{n=1}^{\infty}\Pr(E_n) < \infty$. We apply Lemma 6 to obtain $\lambda_2 < \liminf_{n\to\infty} L(X_1,\ldots,X_n;pn)/(p\sqrt{n})$ with probability one, which completes the proof. □

In order to address the general (i.e. non-uniform) case of the CCTSP, the following lemma is useful:

Lemma 19. *Let $X_1,\ldots,X_n$ be independent uniform samples drawn from a compact region $\mathcal{R}$ with area 1 and let $\mathcal{S}\subset\mathcal{R}$, with $\text{Area}(\mathcal{S}) = q$. Then*

$$\liminf_{n\to\infty}\frac{L(X_1,\ldots,X_n\cap\mathcal{S};pn)}{p\sqrt{n}} \ge \begin{cases}\lambda_2 & \text{if } p\le q\\ \infty & \text{otherwise.}\end{cases}$$

Proof. This is simple: if $p > q$ then the law of large numbers says that $|X_1,\ldots,X_n\cap\mathcal{S}|/n \to q$ with probability one, so $L(X_1,\ldots,X_n\cap\mathcal{S};pn)$ does not exist (we would have to visit more points than are contained in $\mathcal{S}$). On the other hand, if $p\le q$, then we merely observe that

$$L(X_1,\ldots,X_n\cap\mathcal{S};pn) \ge L(X_1,\ldots,X_n;pn)$$

and apply Theorem 18. □

The non-uniform equivalent of Theorem 18 for step functions follows:

Lemma 20. *Let $\phi(x) = \sum_{i=1}^{s} a_i\mathbb{1}(x\in\boxdot_i)$ be a step density function with compact support $\mathcal{R}$ such that $a_1\ge\cdots\ge a_s$ and $a_i\,\text{Area}(\boxdot_i) = 1/s$ for all i (so that $\text{Area}(\mathcal{R}) = 1$). If $\mathscr{Y} = \{Y_i\}$ is a sequence of independent samples from ϕ, $0<p<1$ and $\Pi(x)$ is defined as in Lemma 10, then*

$$\begin{aligned}\lambda_2\iint_{\mathcal{R}}\sqrt{\phi(x)}\mathbb{1}(\Pi(x)\ge p)\,\mathrm{d}x &< \liminf_{n\to\infty}\frac{L(X_1,\ldots,X_n;pn)}{p\sqrt{n}} \\ &\le \limsup_{n\to\infty}\frac{L(X_1,\ldots,X_n;pn)}{p\sqrt{n}} < \mu_2\iint_{\mathcal{R}}\sqrt{\phi(x)}\mathbb{1}(\Pi(x)\ge p)\,\mathrm{d}x\end{aligned}$$

with probability one, where λ_2 and μ_2 are the constants from Theorem 18.

Proof. The upper bound is easy: if we simply take a tour of those points x such that $\Pi(x) \geq p$, then the law of large numbers says that we will visit $\sim pn$ points as $n \to \infty$, and the BHH Theorem (i.e. Theorem 4) We have

$$\begin{aligned} L(Y_1, \ldots, Y_n; m) &= \min_{S \subset \mathscr{Y} : |S| = m} \mathsf{TSP}(S) \\ &= \min_{S \subset \mathscr{Y} : |S| = m} \sum_{i=1}^{s} \mathsf{TSP}(S \cap \boxdot_i) + \mathcal{O}(1) \end{aligned}$$

from Lemma 5. By definition, we also have

$$\min_{S \subset \mathscr{Y} : |S| = m} \sum_{i=1}^{s} \mathsf{TSP}(S \cap \boxdot_i) = \min_{\boldsymbol{q} \in \boldsymbol{Q}} \sum_{i=1}^{s} L(S \cap \boxdot_i; q_i)$$

where $\boldsymbol{q}$ denotes the number of points from each $\boxdot_i$ that are selected, i.e.

$$\boldsymbol{Q} = \left\{ \boldsymbol{q} \in \mathbb{Z}_+^s : \sum_{i=1}^{s} q_i = m \,,\; q_i \leq |\mathscr{Y} \cap \boxdot_i| \; \forall i \right\} .$$

Define Ψ, $\mathscr{Y}'$, and Y_i' as in the proof of Lemma 13, as well as $S' = \Psi(S)$ for $S \subset \mathscr{Y}$. As before, we have

$$\mathsf{TSP}(S \cap \boxdot_i) = \frac{1}{\sqrt{a_i}} \mathsf{TSP}(S' \cap \boxdot_i')$$

for all subsets S. Since the a_i's are decreasing, we can define increasing terms $b_i = 1/\sqrt{a_i}$ and we can also construct c_j's so that

$$\begin{aligned} \min_{\boldsymbol{q} \in \boldsymbol{Q}} \sum_{i=1}^{s} L(S \cap \boxdot_i; q_i) &= \min_{\boldsymbol{q} \in \boldsymbol{Q}} \sum_{i=1}^{s} \frac{1}{\sqrt{a_i}} L(S' \cap \boxdot_i'; q_i) \\ &= \min_{\boldsymbol{q} \in \boldsymbol{Q}} \sum_{i=1}^{s} b_i L(S' \cap \boxdot_i'; q_i) \\ &\geq \min_{\tilde{\boldsymbol{q}} \in \tilde{\boldsymbol{Q}}} \sum_{i=1}^{s} b_i L(S' \cap \boxdot_i'; \tilde{q}_i n) \end{aligned}$$

where $\tilde{\boldsymbol{Q}}$ is a "lower bounding set" of $\boldsymbol{Q}$ defined as follows: fix ϵ and let $\xi(t) = \epsilon \lfloor t/\epsilon \rfloor$, which in particular tells us that $0 \leq t - \xi(t) \leq \epsilon$ for all t. The set $\tilde{\boldsymbol{Q}}$ is the image of $(\lfloor n\xi(q_1/n) \rfloor, \ldots, \lfloor n\xi(q_s/n) \rfloor)$ for all feasible vectors $\boldsymbol{q} \in \boldsymbol{Q}$; in particular, it has the property that for any $\boldsymbol{q} \in \boldsymbol{Q}$, there exists

$\tilde{\boldsymbol{q}} \in \tilde{\boldsymbol{Q}}$ such that $\tilde{\boldsymbol{q}} \leq \boldsymbol{q}$. We have therefore established that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{L(Y_1,\ldots,Y_n;pn)}{\sqrt{n}} &\geq \liminf_{n\to\infty} \frac{1}{\sqrt{n}} \min_{\boldsymbol{q}\in\boldsymbol{Q}} \sum_{i=1}^{s} b_i L(S' \cap \boxdot_i'; q_i n) \\
&\geq \liminf_{n\to\infty} \frac{1}{\sqrt{n}} \min_{\tilde{\boldsymbol{q}}\in\tilde{\boldsymbol{Q}}} \sum_{i=1}^{s} b_i L(S' \cap \boxdot_i'; \tilde{q}_i n) \\
&= \min_{\tilde{\boldsymbol{q}}\in\tilde{\boldsymbol{Q}}} \liminf_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{s} b_i L(S' \cap \boxdot_i'; \tilde{q}_i n) \qquad (14) \\
&\geq \min_{\boldsymbol{t}\in p\Delta^{s-1}} \liminf_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{s} b_i L(S' \cap \boxdot_i'; (t_i-\epsilon) n)
\end{aligned}$$

where the exchange in (14) is permissible because the cardinality of $\tilde{\boldsymbol{Q}}$ remains bounded[1] as $n \to \infty$. We are therefore free to consider the outer optimization problem in $\boldsymbol{t}$, which is

$$\begin{aligned}
&\underset{\boldsymbol{t}}{\text{minimize}}\ \lambda_2 b_i (t_i - \epsilon) \quad s.t. \qquad (15) \\
&\sum_{i=1}^{s} t_i = p \\
&0 \leq t_i \leq 1/s \quad \forall i
\end{aligned}$$

and is obviously minimized by setting $t_1 = \cdots = t_{\lfloor ps \rfloor} = 1/s$ and $t_{\lceil ps \rceil} = p - \lfloor ps \rfloor / s$ (see e.g. exercise 4.8(e) of [3]). We can disregard the $t_{\lceil ps \rceil}$ term for notational convenience and use the fact that $x \in \bigcup_{i=1}^{\lfloor ps \rfloor} \boxdot_i$ if and only if $\Pi(x) \geq \lfloor ps \rfloor / s$ to see that the objective function of (15) is at least

$$\begin{aligned}
\lambda_2 \sum_{i=1}^{\lfloor ps \rfloor} b_i (1/s - \epsilon) &= \frac{\lambda_2}{s} \sum_{i=1}^{\lfloor ps \rfloor} b_i - \epsilon \lambda_2 \sum_{i=1}^{\lfloor ps \rfloor} b_i \geq \frac{\lambda_2}{s} \sum_{i=1}^{\lfloor ps \rfloor} b_i - \epsilon \quad \lambda_2 \sum_{i=1}^{s} b_i \Bigg) \\
&= \int_{\mathcal{R}} \sqrt{\phi(x)}\, \mathbb{1}(\Pi(X) \geq \lfloor ps \rfloor / s)\, \mathrm{d}x - \epsilon \left(\lambda_2 s \int_{\mathcal{R}} \sqrt{\phi(x)}\ \mathrm{d}x \right) \Bigg(\\
&\geq \int_{\mathcal{R}} \sqrt{\phi(x)}\, \mathbb{1}(\Pi(X) \geq p)\, \mathrm{d}x - \epsilon \left(\lambda_2 s \int_{\mathcal{R}} \sqrt{\phi(x)}\ \mathrm{d}x \right) \Bigg(
\end{aligned}$$

which completes the proof, because for fixed s, we can make ϵ as small as we like to reduce the subtracted term. □

The non-uniform equivalent to Theorem 18 follows:

Theorem 21. *Let f, $\mathcal{R}$, and P be as in Lemma 10. We have*

$$\begin{aligned}
\lambda_2 \int_{\mathcal{R}} \sqrt{f(x)}\, \mathbb{1}(P(x) \geq p)\, \mathrm{d}x &< \liminf_{n\to\infty} \frac{L(X_1,\ldots,X_n;pn)}{p\sqrt{n}} \\
&\leq \limsup_{n\to\infty} \frac{L(X_1,\ldots,X_n;pn)}{p\sqrt{n}} < \mu_2 \int_{\mathcal{R}} \sqrt{f(x)}\, \mathbb{1}(P(x) \geq p)\, \mathrm{d}x
\end{aligned}$$

with probability one, where λ_2 and μ_2 are the constants from Theorem 18.

[1] For example, since each entry of $\tilde{\boldsymbol{Q}}$ is of the form $\lfloor i\epsilon n \rfloor$ with $i \in \mathbb{Z}$ satisfying $0 \leq i \leq \lceil 1/\epsilon \rceil$, a crude upper bound is $|\tilde{\boldsymbol{Q}}| \leq (\lceil 1/\epsilon + 1)^s$.

Proof. We omit the proof for brevity because it is almost identical to the proof of Theorem 15, in the sense that we merely show that a step density can approximate a smooth density to any desired precision. □

8 Computational experiments

This section presents the results from two computational experiments. The first two experiments use synthetic data in the unit square, and the third uses road network data to determine if clustering in the GTSP has an impact for reasonable values of n. In order to (heuristically) solve the GTSP, both experiments use the Lin-Kernighan-Helsgaun heuristic adapted for the GTSP described in [16].

8.1 Predicting tour lengths of the generalized TSP

We consider predicting the length of a GTSP tour of uniform samples of points in a unit square. Based on Theorem 11, we predict that the length L of a GTSP tour of n sets of k points each satisfies

$$\lambda_1\sqrt{n/k} \leq L \leq \mu_1\sqrt{n/k}\,.$$

Figure 5a shows that these lengths do indeed lie well within the bounds given. Figure 5b shows the same data, but using a truncated Gaussian distribution in the unit square.

8.2 Predicting tour lengths of the cardinality-constrained TSP

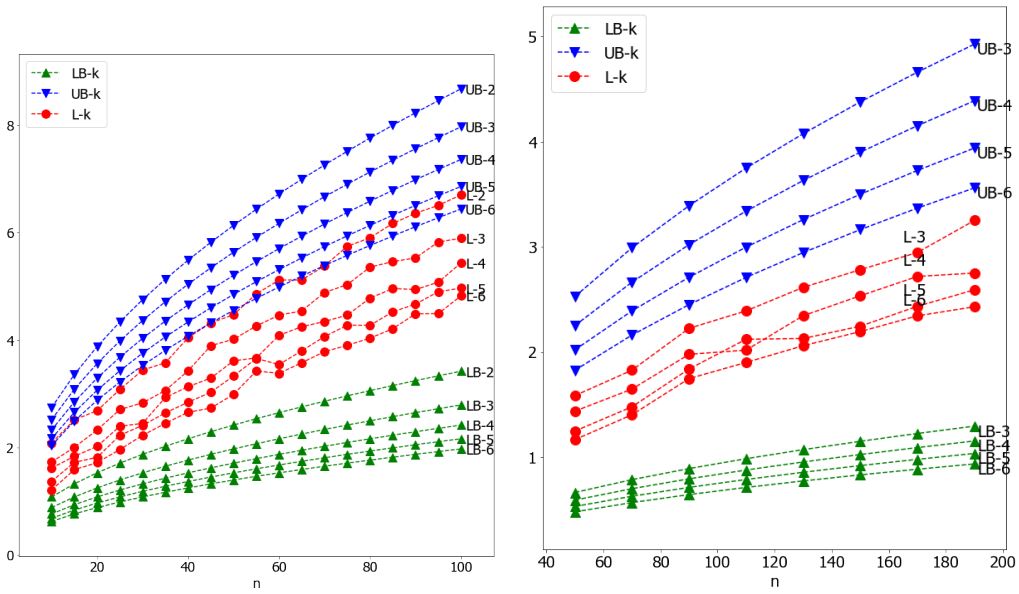
We consider predicting the length of a CCTSP tour of uniform samples of points in a unit square. Based on Theorem 18, we predict that the length L of a CCTSP tour that visits a fraction p of n points satisfies

$$\lambda_2 p\sqrt{n} \leq L \leq \mu_2 p\sqrt{n}\,.$$

Figure 6a shows that these lengths do indeed lie well within the bounds given. Figure 6b shows the same data, but using a truncated Gaussian distribution in the unit square.

8.3 An experiment in a road network

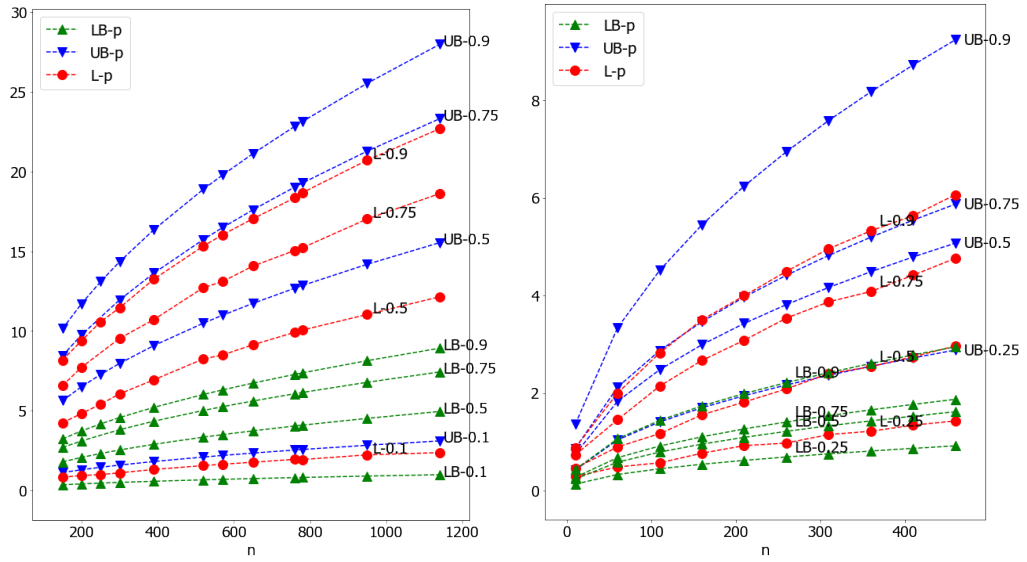
Our second experiment addresses the impact of clustering in the GTSP when distances are measured with respect to a road network, using the Google Distance Matrix API [9]. We again used the Lin-Kernighan-Helsgaun heuristic from [16] for $5 \leq n \leq 100$ and $2 \leq k \leq 4$, and our point sets were sampled uniformly from the centers of the 1856 census blocks of the city of Sunnyvale, California, which are shown in Figure 7. We computed GTSP tours for both the case where all samples are independent as well as the case where each $\mathcal{X}_i$ is clustered: specifically, we let each $\mathcal{X}_i$ consist of samples that are all within one mile of one another (i.e. sampled in a ball of radius 0.5 miles). The tour lengths for these trials, as well as a comparison between the clustered and non-clustered cases, are shown in Figure 8, which indicates that the same basic principles are in effect as in the Euclidean case.



(a) Tour lengths, and upper and lower bounds, for uniformly sampled points in the unit square.

(b) Tour lengths, and upper and lower bounds, for points sampled from a truncated Gaussian distribution in the unit square.

Figure 5: Tour lengths and their bounds. For each diagram, the horizontal axis shows the number of sets $\mathscr{X}_i$, and the vertical axis shows the lengths. The plots show the lower bounds, upper bounds, and true lengths computed, for $k \in \{3, 4, 5, 6\}$.

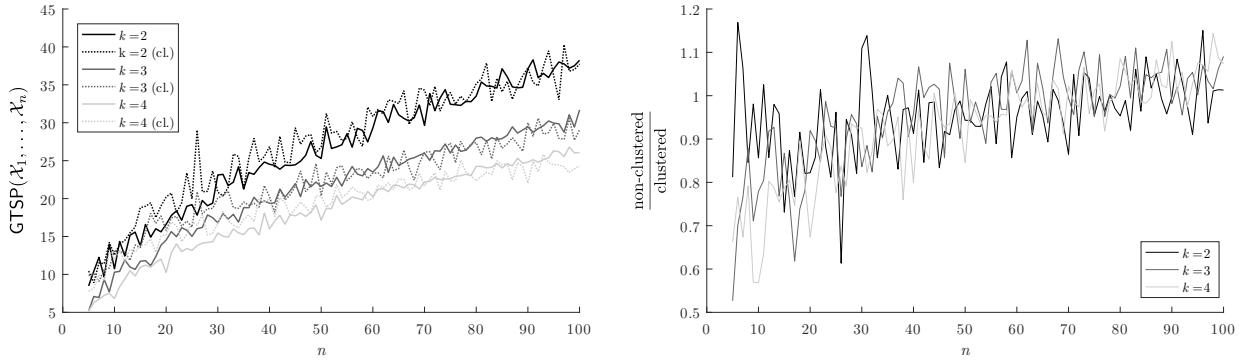


(a) Tour lengths, and upper and lower bounds, for uniformly sampled points in the unit square.

(b) Tour lengths, and upper and lower bounds, for points sampled from a truncated Gaussian distribution in the unit square.

Figure 6: Tour lengths and their bounds. For each diagram, the horizontal axis shows the number of points X_i, and the vertical axis shows the lengths. The plots show the lower bounds, upper bounds, and true lengths computed, for $p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

Figure 7: The centers of census blocks in Sunnyvale, California.



(a) Tour lengths (miles)

(b) Ratios

Figure 8: Figure (a) shows the lengths of the GTSP tours, in miles, of randomly selected centers of census blocks; here the abbreviation (cl.) refers to clustered point sets. In (b), we see that the ratio between clustered and non-clustered lengths indeed approaches 1 as n becomes large, as expected. Also note that the clustered tours are significantly longer than the non-clustered tours when $k = 3, 4$ and n is small.

9 Conclusions

We have proven two theorems that can be used to predict the total length of the solution to a selection routing problem, such as the generalized TSP or the cardinality-constrained TSP. As demonstrated in our computational experiments, these predictions are useful both for Euclidean instances as well as distances on a road network. Further research directions include the integration of temporal constraints such as time windows or capacities on vehicles, and we plan to address them in future work.

References

[1] Henrylab. AL. Record balancing problem-a dynamic programming solution of a generalized traveling salesman problem. *Revue Francaise D Informatique De Recherche Operationnelle*, 3, 1969.

[2] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

[3] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Lawrence D. Burns, Randolph W. Hall, Dennis E. Blumenfeld, and Carlos F. Daganzo. Distribution strategies that minimize transportation and inventory costs. *Operations Research*, 33(3):469–490, 1985.

[5] John Gunnar Carlsson. Dividing a territory among several vehicles. *INFORMS Journal on Computing*, 24(4):565–577, 2012.

[6] John Gunnar Carlsson, Mehdi Behroozi, Raghuveer Devulapalli, and Xiangfei Meng. Household-level economies of scale in transportation. *Operations Research*, 64(6):1372–1387, 2016.

[7] Sudipta Chowdhury, Adindu Emelogu, Mohammad Marufuzzaman, Sarah G. Nurre, and Linkan Bian. Drones for disaster response and relief operations: A continuous approximation model. *International Journal of Production Economics*, 188:167 – 184, 2017.

[8] S. Curtis. Amazon at 15: the technology behind Amazon UK's success. *The Telegraph*, October 15, 2013.

[9] Google Developers. The Google Distance Matrix API. `https://developers.google.com/maps/documentation/distancematrix/intro`, 2015.

[10] L. Few. The shortest path and the shortest road through n points. *Mathematika*, 2:141–144, 1955.

[11] S.R. Finch. *Mathematical Constants*. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 2003.

[12] M. Fischetti, J. J. Salazar G., and P. Toth. A branch-and-cut algorithm for the symmetric generalized traveling salesman problem. *Operations Research*, 45(3):378–394, 1997.

[13] Anna Franceschetti, Ola Jabali, and Gilbert Laporte. Continuous approximation models in freight distribution management. *TOP*, 25(3):413–433, October 2017.

[14] Michel Gendreau, Francois Guertin, Jean-Yves Potvin, and Rene Seguin. Neighborhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries. *Transportation Research Part C: Emerging Technologies*, 14(3):157 – 174, 2006.

[15] Justin C. Goodson, Jeffrey W. Ohlmann, and Barrett W. Thomas. Rollout policies for dynamic solutions to the multivehicle routing problem with stochastic demand and duration limits. *Operations Research*, 61(1):138–154, 2013.

[16] G. Gutin and D. Karapetyan. A memetic algorithm for the generalized traveling salesman problem. *Natural Computing*, 9(1):47–60, 2009.

[17] M. Haimovich and Thomas L. Magnanti. Extremum properties of hexagonal partitioning and the uniform distribution in Euclidean location. *SIAM J. Discrete Math.*, 1:50–64, 1988.

[18] R. W. Hall. Distance approximations for routing manual pickers in a warehouse. *IIE transactions*, 25(4):76–87, 1993.

[19] D. S. Hochbaum. When are NP-hard location problems easy? *Annals of Operations Research*, 1:201–214, 1984.

[20] Dorit Hochbaum and J. Michael Steele. Steinhaus's geometric location problem for random samples in the plane. *Advances in Applied Probability*, 14(1):56–67, 1982.

[21] Michael Huang, Karen R. Smilowitz, and Burcu Balcik. A continuous approximation approach for assessment routing in disaster relief. *Transportation Research Part B: Methodological*, 50:20 – 41, 2013.

[22] Ola Jabali, Michel Gendreau, and Gilbert Laporte. A continuous approximation model for the fleet composition problem. *Transportation Research Part B: Methodological*, 46(10):1591 – 1606, 2012.

[23] D. Karapetyan and G. Gutin. Lin-kernighan heuristic adaptations for the generalized traveling salesman problem. *European Journal of Operational Research*, 208(3):221–232, 2011.

[24] R. Kitamura. Incorporating trip chaining into analysis of destination choice. *Transportation Research Part B: Methodological*, 18(1):67–81, 1984.

[25] T. Lindvall. *The coupling method*. John Wiley, 1992.

[26] C. H. Papadimitriou. Worst-case and probabilistic analysis of a geometric location problem. *SIAM Journal on Computing*, 10:542, 1981.

[27] O. Petrovic, M. J. Harnisch, and T. Puchleitner. Opportunities of mobile communication systems for applications in last-mile logistics. In *Advanced Logistics and Transport (ICALT), 2013 International Conference on*, pages 354–359. IEEE, 2013.

[28] C. Redmond and J. E. Yukich. Limit theorems and rates of convergence for euclidean functionals. *The Annals of Applied Probability*, 4(4):pp. 1057–1073, 1994.

[29] G. Reinelt. Tsplib – a traveling salesman problem library. *ORSA journal on computing*, 3(4):376–384, 1991.

[30] Charles S. Revelle and Gilbert Laporte. The plant location problem: New models and research prospects. *Operations Research*, 44(6):864–874, 1996.

[31] Ulrike Ritzinger, Jakob Puchinger, and Richard F. Hartl. A survey on dynamic and stochastic vehicle routing problems. *International Journal of Production Research*, 54(1):215–231, 2016.

[32] J. P. Saskena. Mathematical model of scheduling clients through welfare agencies. *Journal of the Canadian Operational Research Society*, 1970.

[33] Timothy Law Snyder and J. Michael Steele. Worst-case greedy matchings in the unit d-cube. *Networks*, 20(6):779–800, 1990.

[34] J. M. Steele. Subadditive euclidean functionals and nonlinear growth in geometric probability. *The Annals of Probability*, 9(3):pp. 365–376, 1981.

[35] J. Michael Steele. Growth rates of euclidean minimal spanning trees with power weighted edges. *Ann. Probab.*, 16(4):1767–1787, 10 1988.

[36] J.M. Steele. *Probability Theory and Combinatorial Optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1987.

[37] E. M Stein and R. Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

[38] K. Suh, T. Smith, and M. Linhoff. Leveraging socially networked mobile ICT platforms for the last-mile delivery problem. *Environmental science & technology*, 46(17):9481–9490, 2012.

Data Management Plan

Products of Research

The data that were collected consist of uniformly sampled points in a geographic region as well as lat/lng pairs sampled from Southern California. All origin-destination distances can be computed using OpenStreetMaps, Google Maps, or HERE Maps.

Data Format and Content

There are no files to share; all experiments can be reproduced using only the contents of this paper.

Data Access and Sharing

The general public can access the data from this paper by repeating the experiments that we conducted, which merely require a random number generator.

Reuse and Redistribution

No restrictions to report.

Appendix

A Proof of Lemma 10

In order to reduce our use of subscripts, it will suffice to prove the following:

Claim 22. Let g be a non-negative measurable function with bounded support $\mathcal{S}$ whose level sets have Lebesgue measure zero and assume that $\int_{\mathcal{S}} g$ is rational. For any $\epsilon > 0$, there exists a step function approximation $\psi(x) = \sum_j a_j \mathbb{1}(x \in \boxdot_j)$ of g such that $\int_{\mathcal{S}} |g - \psi| \leq \epsilon$, and

- $\psi(x) = 0$ whenever $x \notin \mathcal{S}$,
- $\inf_{x \in \mathcal{S}} g(x) < a_j < \sup_{x \in \mathcal{S}} g(x)$ for all j,
- a_j and $\text{Area}(\boxdot_j)$ are rational for all j, and
- $\int_{\mathcal{S}} \psi = \int_{\mathcal{S}} g$.

This proves Lemma 10 by substituting $g \mapsto f(x)\mathbb{1}(x \in \mathcal{S}_i)$, $\mathcal{S} \mapsto \mathcal{S}_i$, and $\epsilon \mapsto \epsilon'$.

Proof. For ease of notation, all integrals in this proof are taken over $\mathcal{S}$. We will build a sequence of functions $\sigma \to \tilde{\sigma} \to \varphi \to \tilde{\varphi} \to \psi$. Since simple functions are dense in $L^1(\mathbb{R}^2)$ (see e.g. Theorem 2.4(ii) of [37]), we can approximate g with a simple function $\sigma(x) = \sum_j b_j \mathbb{1}(x \in s_j)$ so that $\int |g - \sigma| \leq \epsilon/8$, where each s_j is measurable. We can assume without loss of generality that $s_j \subset \mathcal{S}$ for all j (otherwise, just set $s_j \mapsto s_j \cap \mathcal{S}$, which can only decrease the distance $\int |g - \sigma|$).

Let $l = \inf_{x \in \mathcal{S}} g(x)$ and $u = \sup_{x \in \mathcal{S}} g(x)$. Certainly, $l\,\text{Area}(\mathcal{S}) < \int g < u\,\text{Area}(\mathcal{S})$ because the level sets of g have Lebesgue measure zero. We can assume without loss of generality that $l \leq b_j \leq u$ for all j, since if $b_j < l$, we can only improve our approximation $\int |g - \varphi|$ by setting $b_j \mapsto l$ (and similarly for u). Let $l' > l$ and $u' < u$ be rational numbers sufficiently close to l' and u' respectively so that $l'\,\text{Area}(\mathcal{S}) < \int g < u'\,\text{Area}(\mathcal{S})$, $(l' - l)\,\text{Area}(\mathcal{S}) \leq \epsilon/8$, and $(u - u')\,\text{Area}(\mathcal{S}) \leq \epsilon/8$. For each component coefficient b_j of σ, define $\tilde{b}_j = \min\{\max\{b_j, l'\}, u'\}$. The function $\tilde{\sigma}(x) = \sum_j \tilde{b}_j \mathbb{1}(x \in s_j)$ satisfies

$$\begin{aligned}\int |\sigma - \tilde{\sigma}| &= \sum_{j: b_j < l'} (l' - b_j)\,\text{Area}(s_j) + \sum_{j: b_j > u'} (b_j - u')\,\text{Area}(s_j) \\ &\leq \sum_{j: b_j < l'} (l' - l)\,\text{Area}(s_j) + \sum_{j: b_j > u'} (u - u')\,\text{Area}(s_j) \\ &\leq \max\{l' - l, u - u'\}\,\text{Area}(\mathcal{S}) \leq \epsilon/8\end{aligned}$$

and therefore $\int |g - \tilde{\sigma}| \leq \int |g - \sigma| + \int |\sigma - \tilde{\sigma}| \leq \epsilon/4$.

We can approximate each piece s_j to arbitrary precision with a finite collection of rectangles $\boxplus_j$, all of which are contained in $\mathcal{S}$; this is just the Lebesgue inner measure. If each collection $\boxplus_j$ is chosen so that $\text{Area}(s_j \setminus \boxplus_j) \leq \epsilon/(8\tilde{b}_j \# \tilde{\sigma})$, where $\#\tilde{\sigma}$ denotes the number of components of $\tilde{\sigma}$, then the resulting *step* function $\varphi = \sum_j \tilde{b}_j \mathbb{1}(x \in \boxplus_j)$ satisfies

$$\int |\tilde{\sigma} - \varphi| = \sum_j \tilde{b}_j\,\text{Area}(s_j \setminus \boxplus_j) \leq \epsilon/8$$

and therefore $\int |g-\varphi| \le \int |g-\tilde{\sigma}| + \int |\tilde{\sigma}-\varphi| \le 3\epsilon/8$.

For ease of notation, we re-index the entries of φ and write $\varphi(x) = \sum_j c_j \mathbb{1}(x \in \boxdot_j)$, where each $\boxdot_j$ is an individual rectangle (the identification of the sets $\boxplus_j$ is no longer of any relevance to us). The penultimate step is to construct a further approximation $\tilde{\varphi}(x) = \sum_j \tilde{c}_j \mathbb{1}(x \in \square_j)$ such that $\tilde{c}_j$ and $\text{Area}(\square_j)$ are rational. This is straightforward; choose $\delta > 0$ sufficiently small so that $\delta \le \epsilon(16 \int \varphi)^{-1}$ and $(1-\delta)c_j > l'$ for all j such that $c_j > l'$. For each j, let $\square_j \subset \boxdot_j$ have rational endpoints with $\text{Area}(\square_j) \ge (1-\delta)\,\text{Area}(\boxdot_j)$, and let $\tilde{c}_j$ be rational with $c_j \ge \tilde{c}_j \ge (1-\delta)c_j$. We have

$$\begin{aligned}
\int |\varphi - \tilde{\varphi}| &= \sum_j c_j \,\text{Area}(\boxdot_j) - \tilde{c}_j \,\text{Area}(\square_j) \\
&\le \sum_j c_j \,\text{Area}(\boxdot_j) - (1-\delta)^2 c_j \,\text{Area}(\boxdot_j) \\
&\le 2\delta \sum_j c_j \,\text{Area}(\boxdot_j) = 2\delta \int \varphi \le \epsilon/8
\end{aligned}$$

and so the step function $\tilde{\varphi}(x) = \sum_j \tilde{c}_j \mathbb{1}(x \in \square_j)$ satisfies $\int |g-\varphi| + \int |\varphi - \tilde{\varphi}| \le \epsilon/2$.

The last step is to define $\psi_t(x) = \min\{\max\{\tilde{\varphi}(x) + t, l'\}, u'\}$ (this simply amounts to shifting $\tilde{\varphi}$ vertically by an amount t but truncating everything below l' or above u'). Clearly, the function $\rho(t) = \int \psi_t$ is continuous and monotonically increasing, and there exist values t^- and t^+ such that $\rho(t^-) < \int g < \rho(t^+)$, and therefore $\rho(t^*) = \int g$ for some t^*. Because $\tilde{c}_j$ and $\text{Area}(\square_j)$ are rational, we know that t^* is rational as well, and therefore the step function $\psi \equiv \psi_{t^*}$ satisfies all four bullet points that we required. The last step is to show that $\int |\tilde{\varphi} - \psi| \le \epsilon/2$, whence $\int |g - \psi| \le \epsilon$. Since $\tilde{\varphi}$ and ψ differ only on a vertical translation, we have

$$\int |\tilde{\varphi} - \psi| = \int (\tilde{\varphi} - \psi) = \int \tilde{\varphi} - \int \psi = \int \tilde{\varphi} - \int g \le \int |\tilde{\varphi} - g| \le \epsilon/2$$

which completes the proof. □