

MIT Open Access Articles

An empirical validation and data#driven extension of continuum approximation approaches for urban route distances

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Merchán, Daniel and Winkenbach, Matthias. 2019. "An empirical validation and data# driven extension of continuum approximation approaches for urban route distances." Networks, 73 (4).

As Published: <http://dx.doi.org/10.1002/net.21874>

Publisher: Wiley

Persistent URL: <https://hdl.handle.net/1721.1/140763>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



Merchan Daniel (Orcid ID: 0000-0003-1822-083X)
 Winkenbach Matthias (Orcid ID: 0000-0002-8237-625X)

An empirical validation and data-driven extension of continuum approximation approaches for urban route distances

Daniel Merchán PhD¹ Matthias Winkenbach PhD¹

¹Center for Transportation and Logistics, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

Correspondence

Daniel Merchán, Center for Transportation and Logistics, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

Email: dmerchan@mit.edu

Funding information

January 28, 2019

Abstract—We introduce a data-driven extension to continuum approximation (CA)-based methods used to predict urban route distances. This extension efficiently incorporates the circuitry of the underlying road network into the approximation method to improve distance predictions in more realistic settings. The proposed extension significantly outperforms traditional methods, which build on the assumption of travel according to the rectilinear distance metric. While only marginally increasing the data collection effort, the proposed extension yields reductions of 26 percent points in mean absolute percentage error compared to traditional approximation methods. The obtained distance estimates are within 5 to 15 percent of near-optimal solutions obtained with a large neighborhood search heuristic, depending on the circuitry of the region and the density of stops. Further, by providing a real-world validation of CA methods, we explore how novel sources of geo-spatial and traffic-related data can be efficiently leveraged to improve the predictive performance of CA methods. The proposed extension is particularly relevant to increase the real-world validity of CA methods applied to large-scale optimization problems in logistics system design and planning within urban areas.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as doi: [10.1002/net.21874](https://doi.org/10.1002/net.21874)

Keywords continuum approximation, traveling salesman problem, urban logistics, circuitry, last-mile delivery, street network analysis

1 Introduction

Routing problems are classic problems in combinatorial optimization, and developing efficient solutions to routing problems has been a topic of interest to researchers and practitioners for decades. Due to their combinatorial nature, even simple variants of these problems rapidly become intractable for mathematically optimal solutions as the problem size increases.

Methods to solve the vehicle routing problem (**VRP**), which generalizes the traveling salesman problem (**TSP**), can be broadly classified according to two schools of thought: discrete methods, and methods based on continuum approximation (**CA**) approaches. Discrete methods aim at obtaining optimum or near-optimum solutions, i.e., a (near-) optimum sequence of points of demand (**PODs**) to be visited by each vehicle. The corresponding optimization objective is typically to minimize total routing cost, distance, or time. While exact (i.e., optimal) solutions can be obtained for problems of up to 200 customers (see, e.g., Poggi and Uchoa [44] and references therein), heuristic and metaheuristic approaches are required to solve larger problem instances.

Classical heuristics, which include the well-known savings algorithm [12] and the sweep algorithms [22], seek to obtain a good and feasible initial solution, which is then typically improved by means of a 3-opt post-optimization step [20]. In general, classical heuristics offer simplicity and speed at the expense of accuracy and flexibility. Metaheuristic approaches, including local search (**LS**) and population search (**PS**) methods, perform more extensive searches of the solution space, resulting in higher quality solutions compared to classical approaches, at the expense of computational efficiency and simplicity [20, 53].

CA-based methods, which are the focus of this paper, aim at efficiently quantifying system performance metrics with minimal data. Generally, these methods rely on simplified analytical forms, including closed form expressions, and on concise summaries of data. In routing problems, for instance, **CA**-based methods have been introduced to approximate the expected distance of a near-optimal delivery route using geometric probability theory to derive simple analytical forms based on the area of the service region and the spatial density of demand realizations, i.e., **PODs** [14]. These methods deliberately ignore detailed distance calculations between **PODs** and do not seek to find detailed specifications of route sequences, which greatly reduces data collection efforts

and computational costs [19]. Instead, these methods aim at deriving system-level performance trade-offs with limited data [15].

The complementary use of discrete and CA-based methods offers interesting opportunities to address large-scale problems of logistics system design and planning inspired by real-world applications [2, 19, 29]. For example, simplified analytical expressions can be helpful to explore the features of these systems and gain insights, before obtaining exact solutions using computationally expensive numerical optimization methods. The potential of combining these two sets of methods is particularly relevant for strategic decisions in logistics system design [2]. For instance, distance and cost approximations have been used to simplify hierarchical distribution problems, such as the location-routing problem (LRP), where routing decisions generally play a subordinate role. Examples of the combined use of discrete and CA-based methods to real-world logistics applications include designing integrated package distribution systems [49], designing and planning large-scale urban logistics networks [56], and planning delivery time-windows in e-commerce settings [1].

Urban last-mile logistics applications hold particular interest to our study. Over the past decade the attention to these applications has grown amongst academics and practitioners, mainly driven by two factors. First, more than half of the global population lives in urban areas since 2012 and this figure is expected to increase further over the next decades, with a particularly rapid speed of urbanization being expected in Africa and Asia. In the Americas and Europe, current urbanization rates are already close to 80% [54]. The continuous growth in urban population size and density is driving ever increasing flows of goods and services into densely populated and congested urban areas. Second, the rapid growth and highly dynamic evolution of e-commerce leads to a massive increase in the amount and fragmentation of local deliveries in urban areas. In 2015 alone, parcel deliveries grew at a 7 – 10% rate in mature markets such as the United States or Germany, and up to 300% in emerging markets [27].

While CA-methods are well suited to support strategic decision-making processes in logistics systems in these increasingly complex urban operational environments, extensions are required to increase the validity of these methods in real-world settings. The extant literature generally assumes idealized road networks to develop and test analytical expressions to predict performance metrics in routing problems. In the specific case of urban applications, travel according to the rectilinear (or L_1) norm is usually assumed. Nevertheless, this assumption does not capture real-world urban road network properties such as non-rectangular layouts, obstacles, road directionality, and other complexities of travel. These complexities significantly affect the efficiency of urban trips, and particularly of *local* trips as shorter trips tend to be more circuitous [30, 33, 34].

To the best of our knowledge, extensions to route distance approximations that incorporate the heterogeneous properties of real-world urban road networks have not yet been introduced. The common use of idealized road network assumptions is partially explained by the usually onerous effort to obtain reliable road network data. Nonetheless, contemporary mapping and navigation technologies offer relatively low-cost and automated options to collect rich traffic and geo-spatial information. The extension presented in this paper targets this void.

The contribution of this paper is twofold. First, we present an empirical real-world validation of traditional **CA**-based methods for predicting urban route distances. Second, we introduce a data-driven extension to account for the circuitry of the underlying urban road network, which ultimately improves the quality of **CA**-based distance predictions. This work emphasizes the importance of leveraging contemporary mapping technologies and novel sources of geo-spatial and traffic-related data to extend traditional approximation methods to address emerging problems in urban logistics and transportation, such as the design and planning of large-scale urban distribution networks. We use the parcel delivery operation of Brazil's largest e-commerce platform, Companhia Digital (**B2W**), in the city of São Paulo to illustrate the practical relevance and impact of our work.

The remainder of this paper is structured as follows. In Section 2, we review the extant literature of **CA**-based methods applied to routing problems in logistics, as well as recent studies on urban road network circuitry. The proposed extension along with an experimental design for an in-depth study of selected urban areas are detailed in Section 3. Our experimental results are summarized, analyzed, and discussed in Section 4. In Section 5, we discuss the practical implications and generalizability of our proposed method, and present our findings from extending our analysis to a city-level study. We conclude with a summary of our findings and recommendations for future research in Section 6.

2 Background

In this section, we review the extant literature on **CA**-based methods applied to **VRPs**. We focus the review on the implications of the distance metrics used and the corresponding parameters. An extended overview of **CA** methods being applied to **TSPs** and **VRPs** is provided in Langevin et al. [29]. The monograph by Daganzo [15] provides a broader overview of **CA** methods and their application to general logistics problems, not limited to routing applications. Franceschetti et al. [19] and Ansari et al. [2] survey the evolution and state-of-the art in **CA**-based approaches applied to general logistics problems. Future research directions in **CA**-based models are also discussed in the survey

paper by Ansari et al. [2]. Smilowitz [48] briefly discusses applications of these methods beyond classic logistics problems.

We exclude from this review a discussion of existing discrete approaches to solving commonly discussed routing problems. We refer the reader to the works of Braekers et al. [7], Montoya-Torres et al. [36] and Toth and Vigo [53] for such reviews.

2.1 Predicting expected route distances

We consider N **PODs**¹ within a geographical region R of area A . We assume a spatial (i.e., two-dimensional) and homogeneous Poisson point process to govern the probabilistic behavior of N : **PODs** are independently and uniformly distributed over R with density parameter γ (**PODs** per unit-area) [31]. We seek to approximate or predict the expected distance D of a (near-)optimal **TSP** tour through N stops in region R . In their seminal paper, Beardwood et al. [5] introduce the following theorem:

$$\lim_{N \rightarrow \infty} E[D(N, A)] = \kappa \sqrt{NA}, \quad (1)$$

where \sqrt{NA} is a non-random function and κ is a proportionality constant that depends upon the distance metric used. As the value of N increases, $E[D(N, A)]$ tends asymptotically to $\kappa \sqrt{NA}$ with probability 1.

Beardwood et al. [5] argue that Equation (1) holds for large areas but might not hold in small areas due to local complications, particularly if N is small. Nevertheless, for fairly convex and compact areas such as circles and squares, this simple asymptotic formula holds well for small values of N , even for $N = 2$ [14, 16]. In general, Equation (1) is robust to different shapes of R and minor deviations from spatial homogeneity of demand within R [14]. This approach can be generalized to non-uniformly distributed points of demand as long as the density is slowly varying over space [10].

From Equation (1), it is straightforward to see that the average inter-stop distance, \bar{d} is given by

$$\bar{d} \approx \frac{\kappa}{\sqrt{\gamma}}. \quad (2)$$

The dimensionless constant κ holds particular interest to our study. Numerous studies have explored suitable values for κ using analytical and empirical approaches for

idealized networks. For the Euclidean (L_2) distance metric, $0.62 \leq \kappa_{L_2} \leq 0.92$ [5] and $\kappa_{L_2} \approx 0.75$ [50]. Given that the distance ratio between rectilinear (L_1) and L_2 metrics for a randomly distributed pair of points is $\frac{4}{\pi} \approx 1.27$, the value of κ for the L_1 metric is given by $\kappa_{L_1} \approx (1.27)(0.75) = 0.95$ [45]. Daganzo [14] analytically derives upper bounds for κ in rectangular regions, specifically $\kappa_{L_2} \leq 0.90$ for the Euclidean metric and $\kappa_{L_1} \leq 1.15$ for the rectilinear metric. Valenzuela and Jones [55] introduce a Held-Karp lower bound $0.708 \leq \kappa_{L_2}$ derived using large problem instances.

Chien [11] extends the study of value ranges for κ_{L_2} in Equation (1) when the shape of the area is not known *a priori* for small problem sizes, $5 \leq N \leq 30$. Using Monte Carlo simulation, regression analysis, and a set of rectangular and sectorial-based shapes, the author finds a best fit value of $\kappa_{L_2} = 0.82$. By comparing CA-based results to exact solutions for all TSP instances tested, his proposed value for κ_{L_2} yields a mean absolute percentage error (MAPE) of 18.6%, whereas $\kappa_{L_2} = 0.75$ yields a MAPE of 23.0%. This result confirms the need for different parameters for lower ranges of N . Chien [11] also suggests a range of minimal-MAPE estimates, $0.73 \leq \kappa_{L_2} \leq 1.29$, depending on the shape of the region.

Kwon et al. [28] further explore the effect of the number of PODs on the value of κ . By examining results for $10 \leq N \leq 80$, they observe that $0.77 \leq \kappa_{L_2} \leq 0.96$. Notice that κ_{L_2} converges to 0.75 as N increases, as predicted by Stein [50]. More recently, building on theoretical arguments on the probabilistic behavior of N , Steinerberger [51] find slight improvements to the bounds for κ_{L_2} originally introduced by Beardwood et al. [5], namely $0.63 \leq \kappa_{L_2} \leq 0.91$.

While the above mentioned authors focus their analyses on the case of the TSP, Daganzo [13] introduces a similar expression for the more general routing problem, the VRP. Assuming a homogeneous vehicle capacity C and an average distance r from the depot to the delivery region, the expectation of the total distance D' for the VRP is given by

$$\lim_{N \rightarrow \infty} E[D'(N, A, C, r)] = 2r \frac{N}{C} + \kappa' \sqrt{NA}. \quad (3)$$

Here, the term $2r \frac{N}{C}$ approximates the so-called ‘line-haul’ portion of the route (i.e., the way from the distribution facility to the start of the delivery route and from the end of the route back to the distribution facility) covered by all $\frac{N}{C}$ vehicles, whereas the term

$\kappa' \sqrt{NA}$ approximates the distance incurred from local distribution. As in Equation 1, the parameter κ' is also a proportionality constant that depends upon the distance metric used. Generally, $\kappa' < \kappa$ (cf., Equation (1)). For the Euclidean metric $\kappa'_{L_2} \approx 0.57$, and, by extension, for the rectilinear metric $\kappa'_{L_1} \approx (1.27)(0.57) \approx 0.73$ [13]. Empirical evidence reveals the superiority of Equation (3) over Equation (1) for experimental settings in which a distance to the depot is included [11, 28]. While not precisely using a **CA** framework due to the deterministic nature of their approach, yet building on the geometric properties of this problem, Haimovich and Rinnooy Kan [25] introduce upper and lower bounds along with a combination of simple closed-form expressions and discrete solution heuristics for this variant of the capacitated **VRP**.

Extensions to the functional form of Equations (1) and (3) are introduced and empirically tested in Chien [11], Kwon et al. [28] and Figliozzi [17]. Overall, these extensions lead to improvements in prediction accuracy, at the expense of requiring more parameters. For instance, Kwon et al. [28] introduce a neural network-based model, which is slightly more accurate but generally less parsimonious.

CA-based methods have also been extended to account for different idealized road network topologies, such as variants of a ring-radial metric [38, 39], and the effect of time windows [18]. Smilowitz and Daganzo [49] present an optimization framework that builds on **CA** of routing costs to design large-scale package distribution systems. Building on this work, Winkenbach et al. [56] introduce an augmented routing cost approximation to account for maximum service time constraints within a mixed-integer linear programming model. This model is used to solve the capacitated two-echelon location-routing problem (**2E-CLRP**) for designing a large-scale urban logistics network. Agatz et al. [1] also build on **CA**-based methods for planning time slots in attended home delivery operations. Nicola et al. [40] introduce a regression-based approach to approximate route distances for the **TSP**, the capacitated vehicle routing problem with time-windows (**CVRP-TW**) and the multi-region multi-depot pickup and delivery problem (**MR-MDPDP**). The authors build on a stepwise regression model to empirically identify key drivers of route distance and cost out of an initially larger set of potentially correlated variables.

Theoretical arguments and empirical evidence suggest that, while line-haul distances are sensitive to the shape of the delivery region, the inter-stop distance (cf., Equation (2)) is mostly sensitive to local conditions, including but not limited to the norm and probabilistic process governing the distributions of **PODs** [13]. The dependence on the norm, and therefore the underlying assumptions about the geometry and topology of the road network, are encoded in the value of α . For urban applications, usually the rectilinear norm is assumed [38, 56].

However, the validity of the classic L_1 norm assumption for urban travel presents limitations [30, 34]. The L_1 norm assumption oversimplifies the underlying road network as it fails to account for features of the road network that affect travel directness, such as topological configuration, obstacles, unidirectional roads, and other complications to travel. These real-world properties of urban road networks entail heterogeneous impacts on the efficiency of local trips [35]. We argue that this oversimplification limits the validity of **CA**-based methods in real-world, large-scale urban logistics applications.

To the best of our knowledge, extensions to account for the effect of real road network features in **CA** have not been introduced. Indeed, the work by Figliozzi [17] is the first to include a real-world case study and suggests to further study the accuracy of VRP distance approximations in cities with different road network configurations. Daganzo [13] and Daganzo [14] conjecture that **TSP** and **VRP** approximations can be extended to other distance metrics simply by multiplying α with the route circuitry factor, but provide no results other than a comparison between the L_2 and L_1 norms. In this study, we provide an empirical real-world validation of this conjecture and discuss how the coefficient α should be calibrated to account for local circuitry conditions.

2.2 Circuitry factors

Circuitry measures the relative detour incurred by vehicles traveling within a road network compared to the straight-line distance between the origin and the destination of their path. A circuitry factor c is thus defined as the ratio between the shortest-path network distance d_c and the Euclidean distance,

$$c = \frac{d_c(p, q)}{d_{L_2}(p, q)}, \quad (4)$$

where a value of c closer to 1 indicates higher levels of network efficiency [4].

In addition to informing network efficiency, circuitry factors can be leveraged to calibrate distance approximations between a pair of points based on the Euclidean norm [34]. Theoretically, if urban travel is assumed to occur only over an isotropic rectilinear

network, then $\bar{c} = \frac{4}{\pi} \approx 1.27$ [31]. Love and Morris [34] empirically found values for c

between 1.16 and 1.28 for selected urban areas in the United States (US), and *circa* 1.35 for rural zones. Similarly, Newell [37] estimate a factor of $\bar{c} = 1.2$ for general urban travel. Ballou et al. [3] introduce inter-city circuitry estimates which range between 1.12 and 2.10, depending upon road network density, connectivity and the presence of geographic obstacles.

In their study of 22 cities in the US, Levinson and El-Geneidy [33] find an average $\bar{c} = 1.18$. Further, they explain city-level road network circuitry based upon a set of network attributes, such as the number of street-to-street and freeway-to-freeway nodes, and road length. Model results suggest that street and freeway length decrease circuitry, i.e. the larger the road length, the higher the likelihood of a direct trip between origin and destination. On the contrary, they observe that the number of street-to-street and freeway-to-freeway nodes increase circuitry. Giacomini and Levinson [21] empirically estimate $\bar{c} = 1.34$ for the 51 most populous metropolitan areas in the United States and find statistically significant evidence of road network efficiency decline between 1990 and 2010 for nearly 70% of the metropolitan areas. Circuitry estimates are weighted by distance traveled in home-work commutes considering trips up to 60 kilometer (**km**). They also observe that circuitry increases inversely proportional to distance, which is also concluded by Levinson and El-Geneidy [33].

Merchán and Winkenbach [35] introduce circuitry factor estimates for local trips. Using the city of São Paulo, Brazil, as the primary case study and considering trips below 5 **km**, they find significant heterogeneities in network circuitry across the city. Local circuitry estimates range between 1.35 and 5.60 with an average value of $\bar{c} \approx 2.50$. Further, the authors introduce a regression model to explain the correlations between local circuitry and road-network properties. In contrast to the findings reported by Levinson and El-Geneidy [33] based on city-level trips covering larger distances (e.g., commuter trips or the line-haul portion of a delivery route), Merchán and Winkenbach [35] observe that local circuitry is negatively correlated with the connectivity of the network and is positively correlated with the length of large-capacity roads (e.g., freeways). Overall, these findings suggests that the effects of road-network properties on the efficiency of urban travel varies depending upon the length of the trip. Results are validated using seven additional case study cities.

3 Proposed Extension and Experimental Design

In this section, we first introduce a proposed extension to account for road network circuitry in **CA**-based methods to predict delivery route distances. Second, we outline the experimental design to test the quality of the proposed extension in real-world settings. To guide the selection of urban areas for the experiment, we conduct a cluster analysis on the properties of the road network to derive classes of city segments having homogeneous road network properties. Next, we introduce the computational experiments to test the proposed extension. The experimental setup described in this paper is inspired by the last-mile delivery operations of **B2W** in the metropolitan area of São Paulo, Brazil.

3.1 Proposed extension

We limit our discussion of our proposed data-driven extension to traditional **CA**-based route distance approximation methods to the **TSP**. However, we conjecture that this extension also applies to distance approximations in the **VRP**, presented in Equation (3), as the term to approximate the local-distribution portion of the route is conceptually and mathematically equivalent to the analytical form of the distance approximation used for the **TSP**, which is presented in Equation (1). We exclude from our analysis the so-called ‘line-haul’ portion of delivery route since, as noted in Section 2, the circuitry of the road network affects predominantly the local-distribution portion.

From Equation (2), the approximate average inter-stop distance for the Euclidean distance metric is given by

$$\overline{d_{L_2}} = \frac{\kappa_{L_2}}{\sqrt{\gamma}}. \quad (5)$$

Then, the average *real* inter-stop distance $\overline{d_c}$ can be approximated by

$$\overline{d_c} \approx \frac{\kappa_c}{\sqrt{\gamma}}, \quad (6)$$

where κ_c is given by

$$\kappa_c = c\kappa_{L_2}. \quad (7)$$

The proposed extension adjusts $\overline{d_{L_2}}$ to account for the local circuitry of the underlying road network encoded in the circuitry factor c . Then, the extended approximation of near-optimal real tour distances, D_c , which preserves the same functional form as in Equation (1), is given by

$$E[D_c] \approx \kappa_c \sqrt{NA}. \quad (8)$$

3.2 Experimental design

We seek to compare the quality of our proposed extension to **CA**-based methods for local route distance estimations with their traditional counterparts in real urban settings. Specifically, we analyze three different approximation models:

Model 1: $D_{L1} = \kappa_{L1} \sqrt{NA}$ (the traditional model)

Model 2: $D_c = \kappa_c \sqrt{NA}$ (the proposed model)

Model 3: $D_f = \kappa_f \sqrt{NA}$ (the benchmark model)

The difference between these three models lies in the value of κ used. Model 1, the traditional model, assumes travel according to the rectilinear distance metric, i.e., the L_1 norm, which encodes the assumption of grid-like travel in urban road segments. Model 2 corresponds to the proposed extension (cf., Equation (8)), which adjusts the Euclidean distance approximation based on the local circuitry of the road network. Model 3 is a benchmark model, for which, using regression analysis, we fit the least-squares estimate κ_f to near-optimal tour distance solutions obtained using a local search heuristic augmented with a large neighborhood search method (**LS-LNS**). The computational experiments to derive these near-optimal solutions are detailed in Section 3.2.2. For the values for κ_{L1} and κ_{L2} , we use the upper bounds for rectangular regions proposed by Daganzo [14] (see Table 1).

3.2.1 Case study for empirical model validation

To guide the selection of city segments to empirically validate the three models presented above, we first generate a classification of city segments based on a cluster analysis of

their dimensional and topological road network properties, to ultimately select segments representing classes sharing similar road network characteristics. Further, the cluster analysis should also help to identify potential outliers, i.e., groups city segments with atypical road network properties which could introduce significant bias to the study.

The core of the metropolitan area of São Paulo, which covers approximately 1,600 square kilometer (km^2), serves as our primary case study to validate and compare the accuracy of the three route length estimation models presented above. We segment this urban area in a grid of square 1- km^2 segments to discretize our geo-spatial data collection and analysis. For each segment $i \in I$, we quantify each dimensional and topological variable (see Tables 2 and 3). OpenStreetMaps (OSM) [52] serves as the source of road network data, which we process using the Python OSMnx module [6].

We consider two sets of variables to characterize the road network of city segments [6]. Dimensional (i.e., metric) measures (see Table 2) inform the spatial distribution of the urban road network, whereas topological measures (see Table 3) characterize its connectivity, centrality and complexity [6]. These properties have been found to be correlated with the level of circuitry in a given segment [35]. To cluster the data, we implement a Gaussian mixture model (GMM) with K -mixture components fitted using an expectation-maximization (EM) algorithm [26]. We inform the value of K (i.e., the number of clusters) by conducting a silhouette score analysis that explores the separation in the dataset as a function of K [46]. In a pre-processing stage, we conduct a principal component analysis (PCA) on the explanatory variables to reduce the dimensionality of the dataset and address multi-collinearity issues among the dimensional and topological variables. The number of principal components (PCs) to use for clustering is defined based on an explained variance threshold ϕ . We implement these un-supervised learning methods in Python using the Scikit-learn module [41].

Results from the GMM-based cluster analysis on the São Paulo metropolitan area, using the six PCs with the highest explanatory power ($\phi = 0.90$), suggest that $K = 3$ clusters (mixtures) provide the largest cluster separation. Thus, we define $K = 3$ classes of urban segments based on dimensional and topological heterogeneities of the road network.

The first class, which covers approximately 20% of the urban area, corresponds to city segments that exhibit a fine-grained road network, but also a large fraction of one-way streets. Furthermore, these segments are often crossed by large-capacity roads such as highways and primary roads. These properties reduce the connectivity of the road network and increase the circuitry of local trips within these segments. As segments from this class are frequently located in city areas with very high levels of ambient population density, they are particularly relevant for urban logistics operations. Thus, we select two

sample segments corresponding to this class. *Segment S1* represents São Paulo's downtown area, one of the most dense and complex areas in the city, due to its commercial, touristic and governmental relevance (see Figure 1a). *Segment S2* (see Figure 1b) covers the surroundings of Av. Paulista, at the intersection of the neighborhoods Jardim Paulista, Bela Vista, and Paraíso. This is one of São Paulo's main financial and commercial zones.

City segments corresponding to the second class, which covers approximately 50% of the urban area, also exhibit a fine-grained road network. However, complications to travel in these segments are lower, making the road network generally well-connected. As a result, the circuitry of the road network of city segments from this class is usually lower compared to segments in the previous class. As a sample of this class, we select *Segment S3* (see Figure 1c), located within the Parque Guaraní neighborhood, which represents a predominantly residential zone with moderate commercial activity.

The third class includes segments with *coarse-grained* networks. These zones typically correspond to atypical or peripheral zones with scant road network development. Due to its lower relevance in terms of potential demand for logistics services, we do not consider any sample segment from this class in this study. We refer the reader to the paper by Merchán and Winkenbach [35] for an extended discussion on the clustering of urban segments based on topological and dimensional properties, and the correlation between these properties and the circuitry of the road network.

Specific demographic and commercial measurements provide further insights into the differences between the three segments selected for model validation (see Table 4). We use LandScan, a global population database developed by the Oak Ridge National Laboratory [9], to obtain ambient population measurements, i.e., average population density over 24 hours. For each segment, we also estimate the average daily density of **PODs**,³, from **B2W**'s transactional records for the year 2015. The delivery density levels observed for the three segments, $\gamma_{S1} = 50$, $\gamma_{S2} = 90$, and $\gamma_{S3} = 10$, constitute a representative range of density levels for **B2W**'s São Paulo operations. Table 4 also includes three selected dimensional and topological variables.

The circuitry factor c holds particular interest to our study as it is the basis for the extension presented in this paper. To quantify c_i per segment $i \in I$, we generate point-to-point trips between T randomly located pairs of points within the segment and obtain road network distances d_c from the Google Distance Matrix (**GDM**) service [23]. We define $T = 180$ based on the sampling method described in [32] to estimate expected values given a specified absolute error of $\epsilon \leq 0.15$. Next, we obtain c_{it} for each trip t according to

Equation (4). Finally, we compute the circuitry factor for segment i , $c_i = \sum_{t=1}^T c_{it} / T$, which we also report in Table 4 for the selected segments S1, S2 and S3.

We observe that denser areas, such as Segments S1 and S2, exhibit more constrained road networks, characterized for instance by a larger proportion of one-way streets. One-way streets, in turn, result in more circuitous local trips, as vehicles are frequently forced to take significant detours through the network, even if the corresponding trip origin and destination are close. This negative effect on the efficiency of local trips is even more profound if, in addition to one-way streets, the road network in these dense segments is characterized by a significant proportion of primary roads, which are generally less accessible, and other complications to local travel.

3.2.2 Computational Experiments

We seek to compare the performance of Models 1, 2 and 3 against near-optimal solutions to the **TSP** for a wide range of values $n_{min} \leq N \leq n_{max}$. Out of the several meta-heuristics available [20], we solve the **TSP** to near-optimality with a local search heuristic augmented with a large neighborhood search method (**LS-LNS**) [47], in an effort to balance accuracy and solution speed. This balance is particularly important given the large number of instances considered in this experiment. Notice that, given the use of real data for **POD** locations and road network distances, the **TSP** is asymmetric and, thus, more computationally expensive to solve. We implement our **LS-LNS** approach using the Google optimization tools routing library [24, 42]. We refer the reader to the manuscript by Pisinger and Ropke [43] for an overview and recent extension of the **LS-LNS** heuristic.

For a set S of randomly generated **PODs** within each segment, which represents the segment's total customer population, we sample N stops to construct a route. The real road network distances for each pair of locations are obtained from the **GDM** service. Furthermore, for each value of N , we solve M randomly generated instances. Inspired by the levels of density (cf., Table 4) and the overall size of the customer population observed for each segment in the case study, we set $n_{min} = 3$, $n_{max} = 100$ and $|S| = 300$. We set $M = 100$ based on an absolute error of $\epsilon \leq 0.2$ km [32]. We also examine the performance of the approximations for three particularly large values of N , namely $N = 150$, $N = 200$, and $N = 250$. In total, we process approximately 10,000 routes per segment. An overview of how our experimental study is conducted is presented in Algorithm 1.

Algorithm 1 Computational experiments in Segments S1, S2 and S3

Input: Geo-coordinates of the corresponding segment
Generate set S by randomly sampling **PODs** based on the defined set size $|S|$
Obtain origin-destination (**OD**) matrix of real distances between each pair of **PODs** in S using the **GDM** service
for $N = n_{min} : n_{max}$ **do**
 for $m = 1 : M$ **do**
 Sample N **PODs** from S to construct a route
 Solve the **TSP** using **LS-LNS** and compute near-optimal tour distance D_m^N
 end for
 Compute the approximated distances for Model 1, D_{L1}^N , Model 2, D_c^N , and Model 3, D_f^N
end for

3.2.3 Estimation of model parameters

Table 5 summarizes the values for $^\circ$ used for each model and for each segment. The value for κ_{L_1} of 1.15 corresponds to the upper bound for the rectilinear distance metric (see Table 1). As defined in Equation (7), the values for $^\circ_c$ are obtained by multiplying κ_{L_2} , the upper bound for the Euclidean distance metric (see Table 1), with c for each segment (see Table 4).

We use ordinary least squares (**OLS**) regression to fit the value of $^\circ_f$ in Model 3 for each segment based on the results of our experiments described in the previous section. The dependent variable is the near-optimal distance of the route D_m^N . Given that the area A is fixed for this experiment, the only independent variable is the number of **PODs**, N . Our regression model includes no intercept term and uses all approximately 10,000 simulated routes from the experiments described in Section 3.2.2 as observations, divided in testing (75%) and training (25%) sets [26]. We implement the following regression model in Python using the Scikit-learn module [41].

Table 6 summarizes our regression results. The relatively high R^2 values for all train and test sets validate the robustness of the functional form $\kappa\sqrt{NA}$ to approximate route distances in real settings. The R^2 values are slightly higher for the least circuitous Segment S3, which is expected because lower circuitry generally leads to lower variability in distances traveled [35]. Finally, we note that the R^2 values we obtain are lower than those reported in Chien [11], Kwon et al. [28], and Figliozzi [17], which are close to 0.99

for the same functional form, $\kappa\sqrt{NA}$. We surmise that this is due to the use of real-world data instead of data obtained from idealized problem instances.

4 Analysis and Discussion of Results

In this section, we report and analyze our experimental results by comparing the route distance estimation performance of Models 1, 2 and 3. Our discussion focuses on the effect of road network circuitry and the number of **PODs**, N , on the quality of the route distance approximations obtained from the three models included in our experiment design, measured by the **MAPE** and the mean percentage error (**MPE**) for each of the models in each of the selected segments.

4.1 Analysis of model performance

We first explore the validity of Model 1 under real urban road network conditions by comparing the distance approximations it produces with the near-optimal results obtained with the **LS-LNS** heuristic [47]. Figure 2a illustrates the results for Segment S2. Each dot represents one route instance solved with **LS-LNS**. The dotted line represents the average near-optimal distance across all instances for each value of N . The solid line illustrates the approximate distances based on Model 1.

We observe that Model 1 consistently and significantly underestimates the near-optimal distances: the **MAPE** ranges between 35% and 44% and is relatively stable for $N \geq 15$ (see Figure 2b). The average **MAPE** for Segment S2 is 43.6%. For Segments S1 and S3, the average **MAPE** is 58.3% and 24.1%, respectively (see Table 7). Overall, Model 1 performs poorly in all three segments suggesting an oversimplification of the underlying road network due to the assumed L_1 metric. We also report the **MPE** to explore over-/under-estimation patterns.

Model 2 yields average **MAPE** values of 12.7%, 8.7% and 10.6% for Segments S1, S2 and S3, respectively. Further, we note that Model 2 tends underestimate distances for Segment S1, the segment with the highest levels of circuitry. Model 3, which slightly over-predicts distances for all segments, yields the lowest errors with average **MAPE** ranging between 7.5% - 8.8%.

In examining Table 7 we make the following additional observations. As expected, Model 3 outperforms Models 1 and 2 in all segments. Still, the performance of Model 2, which is significantly more computationally efficient than Model 3, worsens by no more than 5 **percentage points** in approximation quality compared to Model 3. The observed

MPE values indicate that Model 3 generally over-estimates route distances, while Model 1 under-estimates them due to its inherent over-simplification of the underlying road network. Model 2 under-estimates route distances only for Segment S1, the segment in which the circuitry of the road network is higher. The performance of all three models is best for S3, the segment with lowest circuitry factor c . This result suggests that as the circuitry of the network increases, accurately predicting tours distances becomes more difficult.

We further examine the quality of the approximations across the range of values for the number of **PODs**, N (see Figure 3 and Table 8). We observe that the performance of all three models converges generally for $N > 15$ across all three segments. For segments S1 and S2, the performance of Model 1 worsens as the number of stops increases. A plausible interpretation is that in circuitous segments, travel efficiency worsens as the inter-stop distance shortens (i.e., N increases) since vehicles would have less options to circumvent obstacles and circuitous paths. However, for very large N , we should expect a diminishing effect of circuitry as the routing problem essentially becomes an edge covering problem (also referred to as a route inspection problem).

As expected, our results show that, again, Model 3 generally outperforms Model 2. However, the **MAPEs** for Models 2 and 3 are generally close. For Segment S2, they are nearly identical across all values of N . These are promising results for Model 2, which, as noted above, is significantly less computationally onerous than Model 3.

The performance of **CA**-based methods for small N holds particular interest. Results based on analytical derivations using the Euclidean distance metric suggest that these approximations are very precise, even for $N = 2$, in squared and circular regions [14]. Our results, which are obtained based on squared regions, are less optimistic for small N . Even though for $N < 10$, Model 2 and 3 offer improvements over Model 1, the **MAPE** fluctuates around 20%. It can be argued that for small N , this accuracy issue is less relevant because exact methods could be used instead. However, we argue that it is also important to ensure that tour distance prediction methods perform well for both small and large values of N as substantial variation in the observable delivery density levels can be expected for real-world, large-scale urban logistics systems. We briefly discuss alternative methods of defining a benchmark model that performs well even for smaller values of N in Section 4.1.1 below.

We further observe that Model 2 usually over-estimates distances for Segments S2 and S3 (see negative **MPE** values in Table 7) and tends to under-estimate them for Segment S1. This indicates that the effect of using the average circuitry factor to calibrate Model 2 penalizes the prediction in segments with lower circuitry. Finally, we observe slight

improvements in **MAPE** for Model 1 and 2 for $N > 100$, which provides some evidence to the argument of the diminishing effect of circuitry as $N \rightarrow \infty$, i.e., the **TSP** collapses into a route inspection problem.

In summary, our experimental results suggest that the rectilinear assumption for urban delivery vehicle travel (i.e., Model 1), which is widely used in the extant literature, underestimates the expected near-optimal distance of a local delivery route, on average, by 24 – 59% in **MAPE**, depending on the circuitry of the road network in the area of study. The proposed Model 2 yields a significant improvement in predictive performance, with a **MAPE** of 8 to 12%. The route length estimates obtained from Model 2 are within 5 **percentage points** of the **MAPE** performances of the predictions obtained by fitting an **OLS** regression model to all the data, as in Model 3.

4.1.1 Exploring alternative benchmark models

Up until now, we have been using Model 3, an **OLS** regression based on the functional form $\kappa\sqrt{NA}$ defined in Equation (1), as the benchmark model for our study. To validate our findings, we are interested in analyzing the performance of alternative tour distance prediction methods which are independent of this functional form, such as a random forest regression [8]. The goal of this analysis is to explore whether random forests can yield better performance results, particularly for small N scenarios. To fit the random forest model we use the *Scikit-learn* Python module [41].

We compare the predictive performance of Model 3 to the one of a random forest regression model for Segments S1, S2 and S3. In Table 9, we report the obtained performance results i) for all values of N , and ii) for all $N < 15$ (i.e., for small N cases). Our results indicate that the random forest model generally outperforms Model 3. While these performance improvements are relatively small (changes in R^2 of < 0.01) when considering all values of N , the observed improvements in R^2 range between 0.02 and 0.06 when considering only cases with $N < 15$, depending on the segment.

While a detailed comparison of **OLS**-based and random forest regressions falls outside the scope of this study, we argue that the application of random forest and other *state-of-the-art* machine learning methods to predict route distances should be further explored. Our analysis indicates that random forest models could offer predictive performance improvements over existing models. Furthermore, they may be better suited to conduct more complex predictions, such as estimating route distances and travel times based on multiple predictors.

5 Practical Implications

As discussed in Section 1 and 2, a **CA**-based method is appealing for planners as it allows for an efficient analysis of key logistics system design trade-offs based on limited data. The extension proposed in this paper can help urban logistics planners improve distance approximations at relatively low additional data collection cost. Contemporary mapping technologies and traffic databases offer an opportunity to efficiently estimate the circuitry factor of a delivery region, which can then be used to better calibrate traditional models.

Compared to Model 1, the proposed extension in Model 2 yields important error reductions: The observed **MAPE** reduces from 58.25% to 12.65% for Segment S1 ($\Delta_{S1} = 45.60$ percentage points), from 43.56% to 8.72% for Segment S2 ($\Delta_{S2} = 34.84$ percentage points), and from 24.14% to 10.62% for Segment S3 ($\Delta_{S3} = 13.52$ percentage points). As discussed in Section 4, the performance of both models depends on the level of circuitry and the number of stops N , corresponding to the number of **PODs** to be served. In this section, we extend our analysis by considering a larger set of city segments in order to validate and generalize our findings.

For practical applications, the overall error reductions across an entire urban area is a measure of interest. To determine the city-wide effect of introducing a circuitry-based extension to **CA**-based methods to predict urban route distances, we consider again the core of the metropolitan area of São Paulo introduced in Section 3.2.1. Within this urban area, **B2W** serves, on average, nearly 15, 000 orders on any given day. We focus our attention on the segments $i \in I$ corresponding to the first two classes of city segments as suggested by the cluster analysis presented in Section 3.2.1. These segments cover about 70% of the urban area and concentrate close to 90% of the demand served by **B2W**. We characterize each segment $i \in I$ by the corresponding circuitry factor c_i and the average daily density of demand 3_i served by **B2W** (see Figure 4). To compute c_i , we follow the calculations presented in Section 3.2.1. The value 3_i is obtained by processing a year-worth of orders served by **B2W** in 2015. The average density of deliveries per segment is $\bar{\gamma} = 9.0$ **PODs/km²**, whereas the average circuitry factor amounts to $\bar{c} = 2.5$.

We project the overall city-level error reduction from Model 1 to Model 2 using an experimental design similar to the one described in Section 3. The main difference relates to the number of segments i and the examined values of N . In Section 3, we compare the performance of the three models for a wide range of values of N in three selected segments. For the city-level analysis described in this section, we analyze a much larger set of segments $I' \in I$. Specifically, we consider a subset I' with $|I'| = 300$ city

segments, which are selected using a stratified sampling approach to generate an equal fraction of segments corresponding to each of the two classes considered.

Furthermore, unlike the experiment described in Section 3, in which we considered a large range of values of N per segment, for the city-level study described in this section we only consider the performance of Models 1 and 2 for specific values of N per segment $i \in I'$, based on its corresponding γ_i (see Algorithm 2). Specifically, using the nearest integer $j \gamma_i n$ as a proxy of N per segment, we predict the corresponding tour distance using Models 1 and 2, and evaluate the accuracy of these predictions against the near-optimal solutions obtained with the **LS-LNS** [47] for the corresponding number of **PODs**, N . We use the same sample size $M=100$ (corresponding to an error of $\epsilon \leq 0.2$ km), as in Section 3.

Algorithm 2 City-level computational experiment

```

for  $i \in I'$  do
  Input: Geo-coordinates of segment  $i$ 
   $N_i = j \gamma_i A_i n$ 
  for  $m = 1 : M$  do
    Sample  $N_i$  PODs to construct a route
    Obtain OD matrix of real distances between each pair of PODs using the GDM service
    Solve the TSP using LS-LNS and compute near-optimal tour distance  $D_m^i$ 
  end for
  Compute the approximated distances for Model 1,  $D_{L1}^i$ , and Model 2,  $D_c^i$ 
end for

```

Our experimental results suggest that the **MAPEs** of all segments $i \in I'$ for Model 1 range between 20 and 80% (see Figure 5). Higher error ranges are observed for segments that also exhibit higher levels of road network circuitry. Interestingly, the number of served **PODs**, N , does not seem to impact the performance of Model 1. On the other hand, the corresponding **MAPEs** for the proposed Model 2 (see Figure 6) ranges between 10 and 30% for most segments. Note, however, that performance decreases significantly for $N < 10$ and for segments with $c > 3$. These findings are consistent with the in-depth analyses for the three selected segments described in Section 4.

Finally, we quantify the error reduction γ_i obtained from Model 2 for each segment and calculate the mean error reduction $\bar{\Delta}$ across all segments. The average **MAPEs** over all segments $i \in I'$ for Model 1 and Model 2 amount to is 43.09% and 17.28%,

respectively, which yields a projected error reduction across the entire urban area of $\bar{\Delta} \approx 25.81$ percentage points.

Even though our proposed extension reduces the projected error across the city by a nearly 60 % (from 43.09 % to 17.28 %), the average projected **MAPE** of Model 2 remains around 17% . The primary reason for this result relates to the large number of city segments with relatively low delivery density levels, ³. For nearly 65% of all city segments, the density of stops is $\gamma \leq 10$ (cf., Figure 4). As discussed in Section 4, the quality of **CA**-based route distance approximations decreases for low numbers of stops, N . Thus, further research should address this limitation.

6 Conclusion

In this paper, we introduce a data-driven extension to traditional continuum approximation (**CA**)-based methods used to predict near-optimal route distances. The proposed extension efficiently incorporates the circuitry of the underlying road network into the approximation, which yields significant accuracy improvements compared to approximations based on the traditional rectilinear distance assumption for urban travel. For demand areas with more than ten stops, $N > 10$, the mean absolute percentage error (**MAPE**) of the route distance approximations obtained from our proposed extension ranges between 5 and 15%, depending on the circuitry of the underlying road network, compared to near-optimum solutions obtained by using a **LS-LNS**. At the same time, the error performance lies within 5 percentage points of the route distance estimates obtained from a benchmark model fitted using ordinary least squares (**OLS**) regression. Overall, our empirical results indicate that **CA** methods to estimate urban route lengths are robust for a wide range of problem sizes under real road network conditions, subject to proper fitting of the proportionality parameter θ . Our proposed extension provides a efficient alternative to improve the calibration of this parameter.

Moreover, to the best of our knowledge, this paper is the first to provide a real-world validation of traditional **CA**-based route distance approximation methods. We show that oversimplifications of the underlying road network, such as the commonly assumed L_1 metric, result in prediction errors ranging between 20 and 40% for less circuitous areas, and between 40 and 80% for more circuitous areas. These findings are based on a large, city-wide, real-world case study covering approximately 300 km² of the metropolitan area of São Paulo, Brazil, and considering different levels of demand intensity and road network circuitry.

The proposed method to incorporate real road network complications in route length estimations yields an average **MAPE** of approximately 17%, compared to a 43% **MAPE**

obtained with traditional methods. The performance of the proposed extension declines for routes with a lower number of stops to be served (i.e., $N \leq 10$). One modeling alternative could entail, instead of using a city segmentation approach with homogeneous demand area sizes and heterogeneous numbers of stops per area, using heterogeneous demand area sizes which guarantee a minimum level of demand intensity (i.e., $N \geq 10$) per area. However, alternative distance prediction approaches, independent of the classic functional form for density-based route length approximations should also be explored. In this paper, we conduct an initial analysis with a random forest model, which yields promising prediction error improvements between 5 and 10 percentage points for small N .

In this paper, we focus our analysis on the functional form of the distance approximation for the traveling salesman problem (**TSP**) introduced in Beardwood et al. [5]. However, our results are easily transferable to the functional forms of the route distance approximations for the vehicle routing problem (**VRP**) introduced Daganzo [13] and subsequent extensions. Our work also confirms that contemporary mapping technologies and new sources of traffic-related and geo-spatial data can be efficiently leveraged to increase the real-world validity of classic planning methods applied to new logistics problems, such as the design and planning of large-scale urban distribution networks.

We argue that the findings presented in this paper are generalizable to urban areas around the world. Circuity has been found to be related to several dimensional and topological properties of the road network, such as the presence one-way streets or the level of connectivity among its nodes [35]. While the combined effect of these dimensional and topological properties yields different levels of circuity in urban segments within and across cities, the results based on the city-wide experiment introduced in Section 5 suggest that our proposed method is robust to a wide range of levels of circuity, and is thus applicable to a broad spectrum of possible real-world urban road network configurations.

Future work should address the accuracy limitations of **CA**-based route length approximation methods for problem instances with a low number of stops, N . Furthermore, driven by current trends in last-mile logistics and the growing congestion challenges in cities, planners might be also interested in improving the prediction of time-based performance metrics, which could be achieved by incorporating a heterogeneous temporal dimension to the Poisson point process governing the spatial distribution of demand. Alternative prediction methods such as random forests should also be explored as they offer enhanced modeling flexibility and potentially higher predictive performance that can be leveraged to predict a variety of route performance metrics.

7 Acknowledgments

The authors thank **B2W** for contributing to this research through the provision of detailed, real-world data.

References

- [1] Agatz N, Campbell A, Fleischmann M, Savelsbergh M. Time Slot Management in Attended Home Delivery. *Transportation Science* 2011;45(3):435–449.
- [2] Ansari S, Ba_dere M, Li X, Ouyang Y, Smilowitz K. Advancements in continuous approximation models for logistics and transportation systems: 19962016. *Transportation Research Part B: Methodological* 2018;107:229–252.
- [3] Ballou RH, Rahardja H, Sakai N. Selected country circuitry factors for road travel distance estimation. *Transportation Research Part A: Policy and Practice* 2002 11;36(9):843–848.
- [4] Barthélemy M. Spatial networks. *Physics Reports* 2011;499(1-3):1–101.
- [5] Beardwood J, Halton JH, Hammersley JM. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society* 1959;55(04):299.
- [6] Boeing G. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 2017;65:126–139.
- [7] Braekers K, Ramaekers K, Van Nieuwenhuyse I. The vehicle routing problem: State of the art classification and review. *Computers and Industrial Engineering* 2016;99:300–313.
- [8] Breiman L. Random forests. *Machine Learning* 2001;45(1):5–32.
- [9] Bright EA, Coleman PR, Rose AN, Urban ML, LandScan. Oak Ridge, TN: Oak Ridge National Laboratory; 2015. <http://www.ornl.gov/landscan/>.
- [10] Campbell JF. Comments on: Continuous approximation models in freight distribution management. *Top* 2017;25(3):434–437.

- [11] Chien TW. Operational estimators for the length of a traveling salesman tour. *Computers and Operations Research* 1992;19(6):469–478.
- [12] Clarke G, Wright JW. Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research* 1964;12(4):568–581.
- [13] Daganzo CF. The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application. *Transportation Science* 1984 11;18(4):331–350.
- [14] Daganzo CF. The length of tours in zones of different shapes. *Transportation Research Part B: Methodological* 1984 1;18(2):135–145.
- [15] Daganzo CF. *Logistics Systems Analysis*. New York: Springer-Verlag Berlin Heidelberg; 2005.
- [16] Eilon S, Watson-Gandy CDT, Christofides N. *Distribution Management: Mathematical Modelling and Practical Analysis*. New York: Hafner; 1971.
- [17] Figliozzi M. Planning Approximations to the Average Length of Vehicle Routing Problems with Varying Customer Demands and Routing Constraints. *Transportation Research Record: Journal of the Transportation Research Board* 2008;2089:1–8.
- [18] Figliozzi M. Planning approximations to the average length of vehicle routing problems with time window constraints. *Transportation Research Part B: Methodological* 2009;43(4):438–447.
- [19] Franceschetti A, Jabali O, Laporte G. Continuous approximation models in freight distribution management. *Top* 2017;25(3):413–433.
- [20] Gendreau M, Potvin JY. *Handbook of Metaheuristics*, vol. 157. 2nd ed. New York: Springer; 2010.
- [21] Giacomini DJ, Levinson DM. Road network circuitry in metropolitan areas. *Environment and Planning B: Planning and Design* 2015;42(6):1040–1053.
- [22] Gillett BE, Miller LR. A heuristic algorithm for the vehicle-dispatch problem. *Operations Research* 1974;22(2):340–349.

- [23] Google, Google Distance Matrix API; 2017.
<https://developers.google.com/maps/documentation/distance-matrix/>.
- [24] Google, Google Optimization Tools; 2017.
<https://developers.google.com/optimization/>.
- [25] Haimovich M, Rinnooy Kan AHG. Bounds and Heuristics for Capacitated Routing Problems. *Mathematics of Operations Research* 1985;10(4):527–542.
- [26] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer; 2009.
- [27] Joeress M, Schröder J, Neuhaus F, Klink C, Mann F. *Parcel delivery: The future of last mile*. McKinsey&Company; 2016.
- [28] Kwon O, Golden B, Wasil E. Estimating the length of the optimal TSP tour: An empirical study using regression and neural networks. *Computers and Operations Research* 1995;22(10):1039–1046.
- [29] Langevin A, Mbaraga P, Campbell JF. Continuous approximation models in freight distribution: An overview. *Transportation Research Part B: Methodological* 1996;30(3 PART B):163–188.
- [30] Larson RC, Li VOK. Finding minimum rectilinear distance paths in the presence of barriers. *Networks* 1981;11(3):285–304.
- [31] Larson RC, Odoni A. *Urban Operations Research*. 1st ed. Upper Saddle River, New Jersey: Prentice Hall; 1981.
- [32] Law AM, Kelton D. *Simulation Modeling and Analysis*. 3rd ed. New York: McGraw-Hill; 2000.
- [33] Levinson D, El-Geneidy A. The minimum circuitry frontier and the journey to work. *Regional Science and Urban Economics* 2009;39(6):732–738.
- [34] Love RF, Morris JG. *Mathematical Models of Road Travel Distances*. *Management Science* 1979;25(2):130–139.
- [35] Merchán D, Winkenbach M. *Quantifying and analyzing the impact of urban road networks on the efficiency of local trips*; 2018.

- [36] Montoya-Torres JR, López Franco J, Nieto Isaza S, Felizzola Jiménez H, Herazo-Padilla N. A literature review on the vehicle routing problem with multiple depots. *Computers & Industrial Engineering* 2015;79:115–129.
- [37] Newell GF. *Traffic Flow on Transportation Networks*. Cambridge, MA: The MIT Press; 1980.
- [38] Newell GF, Daganzo CF. Design of multiple vehicle delivery tours - II: Other metrics. *Transportation Research Part B: Methodological* 1986;20(5):345–363.
- [39] Newell GF, Daganzo CF. Design of multiple-vehicle delivery tours-I a ring-radial network. *Transportation Research Part B: Methodological* 1986;20(5):345–363.
- [40] Nicola D, Vetschera R, Dragomir A. Total distance approximations for routing solutions. *Computers and Operations Research* 2019;102:67–74.
- [41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2012;12:2825–2830.
- [42] Perron L. Operations Research and Constraint Programming at Google. In: Lee J, editor. *Principles and Practice of Constraint Programming CP 2011* Berlin, Heidelberg: Springer; 2011..
- [43] Pisinger D, Ropke S. Large Neighborhood Search. In: Gendreau M, Potvin JY, editors. *Handbook of Metaheuristics*, 2nd ed. New York: Springer; 2010.p. 399–420.
- [44] Poggi M, Uchoa E. New Exact Algorithms for the Capacitated Vehicle Routing Problem. In: Toth P, Vigo D, editors. *Vehicle Routing Problems, Methods and Applications*, 2 ed.; 2014.p. 59–86.
- [45] Robusté F, Daganzo CF, Souleyrette RR. Implementing Vehicle Routing Models. *Transportation Research Part B: Methodological* 1990;24(4):263–286.
- [46] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987;20(C):53–65.
- [47] Shaw P. Using Constraint Programming and Local Search Methods to Solve Vehicle Routing Problems. In: Maher M, Puget JF, editors. *Principles and Practice of Constraint*

Programming CP98, vol. 1520 of Lecture Notes in Computer Science Berlin, Heidelberg: Springer; 1998. p. 417–431.

[48] Smilowitz K. Comments on: Continuous approximation models in freight distribution management. *Top* 2017;25(3):440–442.

[49] Smilowitz K, Daganzo CF. Continuum approximation techniques for the design of integrated package distribution systems. *Networks* 2007;50(3):183–196.

[50] Stein DM. An Asymptotic, Probabilistic Analysis of a Routing Problem. *Mathematics of Operations Research* 1978;3(2):89–102.

[51] Steinerberger S. New bounds for the traveling salesman constant. *Advances in Applied Probability* 2015;47(1):27–36.

[52] The OpenStreetMap Foundation, OpenStreetMap; 2017.
<http://www.openstreetmap.org/>.

[53] Toth P, Vigo D, editors. *Vehicle Routing Problems, Methods and Applications*. 2nd ed. Society for Industrial and Applied Mathematics, Mathematical Optimization Society; 2014.

[54] United Nations Population Division. *World Urbanization Prospects, the 2014 Revision*. United Nations; 2014.

[55] Valenzuela CL, Jones AJ. Estimating the Held-Karp Lower-Bound for the Geometric TSP. *European Journal of Operational Research* 1997;102(1):157–175.

[56] Winkenbach M, Kleindorfer PR, Spinler S. Enabling Urban Logistics Services at La Poste through Multi-Echelon Location-Routing. *Transportation Science* 2016;50(2):520–540.

(a) Segment S1 (b) Segment S2 (c) Segment S3

Fig. 1 Selected 1-km² segments exhibiting different dimensional and topological properties. For instance, the fraction of directionally-constrained streets (red links) varies significantly across segments.

(a) Near-optimal route distances (b) MAPE

Fig. 2 Comparison of near-optimal tour distances D between the solutions obtained with LS-LNS heuristic and the approximation based on Model 1, for Segment S2.

(a) Segment S1 (b) Segment S2 (c) Segment S3

Fig. 3 MAPE for Models 1, 2 and 3 as a function of N

Fig. 4 Segment density 3 versus average circuitry \bar{c} for 1600 1-km² segments in São Paulo.

Fig. 5 MAPE for Model 1 for 300 city segments

Fig. 6 MAPE for Model 2 for 300 city segments

Table 1 Upper bound values for κ for the Rectilinear and Euclidean distance metrics

κ	Distance Metric
κ_{L_1}	Rectilinear or L_1 norm 1.15
κ_{L_2}	Euclidean or L_2 norm 0.90

Table 2 Segment-level metric variables

Variables	Description
Intersection density (/km ²)	Number of road intersections
Highway length (km)	Total length of highway roads
Primary road length (km)	Total length of primary roads
Street length (km)	Total length of non-highway and non-primary roads
One-way fraction (%)	Fraction of the total street length having a directional constraint (i.e., one-way streets)

Variables	Description
Avg. road-link length (km)	Mean road-link length, including streets, primary roads and highways
Definitions adapted from Boeing [6]	

Table 3 Segment-level topological variables

Variables	Description
Node connectivity	Average number of nodes to remove to disconnect a non-adjacent pair of random nodes
Node degree	Number of edges (streets) emanating from each node, averaged over all nodes
Neighborhood Degree	Average node degree of a node's neighbors, averaged over all nodes
Betweenness centrality	Number of shortest paths that pass through a node, averaged over all nodes
Closeness centrality	Reciprocal of the sum of the distance from the node to all other nodes, averaged over all nodes
Degree centrality	Fractions of nodes that each node is connected to, averaged over all nodes
Definitions adapted from Boeing [6]	

Table 4 Comparison of segments based on demographic, dimensional and topological variables.

Variables	Segment S1	Segment S2	Segment S3
Ambient population (inh.)	28,700	18,200	12,700
Average daily density ³ (POD/km ²)	50	90	10
One-way street length (%)	94 %	73 %	9 %
Highway & primary road length (km)	3.40	2.70	0.00
Avg. node connectivity	0.99	1.01	1.83
Circuitry factor <i>c</i>	2.76	2.34	1.82

Table 5 θ values for each model per each segment

	S1	S2	S3
κ_{L_1}	1.15	1.15	1.15
θ_c	2.48	2.10	1.64
θ_f	2.78	2.07	1.52

Table 6 Regression results for Model 3 per segment

Segment	θ_f	Std. Err.	<i>p</i> -value	R^2 Train	R^2 Test	No. Observations
S1	2.78	0.005	0.00	0.93	0.93	9,960
S2	2.07	0.004	0.00	0.93	0.93	10,090
S3	1.52	0.003	0.00	0.94	0.94	10,080

Table 7 Overall MAPE per segment

Segment	Model 1 κ_{L_1}		Model 2 σ_c		Model 3 σ_f	
	MAPE (%)	MPE (%)	MAPE (%)	MPE (%)	MAPE (%)	MPE (%)
S1	58.28	58.28	12.65	9.55	8.80	-1.34
S2	43.56	43.38	8.72	-3.02	8.53	-1.54
S3	24.14	23.03	10.62	-8.21	7.51	-0.17

Table 8 MAPE per segment for intervals of N

N	Segment	Model 1 κ_{L_1}		Model 2 σ_c		Model 3 σ_f	
		MAPE (%)	MPE (%)	MAPE (%)	MPE (%)	MAPE (%)	MPE (%)
≤ 10	S1	51.48	51.41	19.37	-4.95	23.61	-17.60
	S2	36.29	35.86	21.58	-16.96	20.72	-15.53
	S3	25.67	24.36	16.06	-7.74	14.64	0.26
11–20	S1	57.68	57.69	13.39	7.78	11.32	-3.33
	S2	41.69	41.67	12.27	-6.36	11.81	-5.06
	S3	25.28	25.20	10.72	-6.54	9.01	1.38
21–40	S1	58.85	58.85	12.89	11.12	8.48	0.40
	S2	44.13	44.13	8.75	-1.87	8.63	-0.63
	S3	24.42	24.42	10.31	-7.65	7.50	0.35
41–60	S1	59.19	59.19	12.27	11.84	7.08	1.22
	S2	45.23	45.23	7.28	0.12	7.39	1.34
	S3	23.92	23.92	10.13	-8.37	6.87	-0.32
61–80	S1	59.10	59.10	12.27	11.62	6.97	0.98
	S2	44.68	44.68	6.19	-0.87	6.17	0.36
	S3	23.88	23.88	9.71	-8.42	6.36	-0.37
81–100	S1	58.75	58.75	11.38	10.89	5.88	0.15

		Model 1		Model 2 σ_c		Model 3 σ_f	
		κ_{L_1}					
> 100	S2	44.37	44.37	5.77	-1.45	5.68	-0.21
	S3	23.74	23.74	9.67	-8.62	5.94	-0.55
	S1	54.58	54.58	4.81	1.88	10.15	-9.94
	S2	40.19	40.19	9.36	-9.05	8.27	-7.72
	S3	20.13	20.13	13.83	-13.76	6.37	-5.31

Table 9 Performance comparison of Model 3 and a random forest regression model

Segment	Model 3 σ_f			Random Forest		
	R^2 Train	R^2 Test	R^2 Test N <15	R^2 Train	R^2 Test	R^2 Test N <15
S1	0.93	0.93	0.66	0.94	0.94	0.72
S2	0.93	0.93	0.64	0.94	0.94	0.69
S3	0.94	0.94	0.67	0.94	0.94	0.69

Footnote

¹the terms **POD** and customer stop will be used interchangeably in this paper