Planning Approximations to the Average Length of Vehicle Routing Problems with Varying Customer Demands and Routing Constraints

Miguel Andres Figliozzi

This paper studies approximations to the average length of vehicle routing problems (VRPs). The approximations are valuable for strategic and planning analysis of transportation and logistics problems. The focus is on VRPs with varying numbers of customers, demands, and locations. This modeling environment can be used in transport and logistics models that deal with a distribution center serving an area with daily variations in demand. The routes are calculated daily on the basis of what freight is available. New approximations and experimental settings are introduced. Average distance traveled is estimated as a function of the number of customers served and the number of routes needed. Approximations are tested in instances with different customer spatial distributions, demand levels, numbers of customers, and time windows. Regression results indicate that the proposed approximations can reasonably predict the average length of VRPs in randomly generated problems and real urban networks.

In many logistics problems it is necessary to estimate the distance that a fleet of vehicles travel to meet a set of customer demands. Traveled distance is not only an important element of carriers' variable costs but also a key input in tactical and strategic models to solve problems such as facility location, fleet sizing, and network design. In particular, this research focuses on the ubiquitous case of a depot or distribution center (DC) serving up to N potential customers in the DC's delivery region. In many practical situations, not all potential customers request a visit on the same day. The number of customers served per day, n, may be significantly smaller than N. There may also be a significant variation in the number of customers visited per day of the week (e.g., early weekdays versus weekends). The amount to be delivered or picked up may also vary on a daily basis (e.g., from one to several pallets), as might other requirements such as time window constraints. The daily customer demand is known a night in advance; hence, each daily route and sequence of customers depends on what freight is available on a particular day for delivery or pickup. Although there is variability in the amount and characteristics of the day-to-day demand, the vehicle routing problem (VRP) analyzed in this paper is neither dynamic nor stochastic since all the information related to the customers' demands is known before the vehicles leave the depot or DC. The routes are designed daily, and the number of routes and distance needed depend on the available freight.

Despite the growing implementation of customer-responsive and made-to-order supply chains, the impact of variations on the number of customer requests and demands on average VRP distance traveled has not yet been studied in the literature. All experimental studies have focused on the approximation of the length of specific traveling salesman problem (TSP) or VRP instances (i.e., given a set of customer demands known a priori, how well a given formula approximates the real distance of one specific instance). This research has a different objective: given N potential customers and a variable customer demand (locations, demands, time windows, etc.) in a service area, how well a given formula approximates the average distance of VRP solutions for different levels of n and routing constraints.

LITERATURE REVIEW

There is an extensive body of TSP- and VRP-related literature in operations research and transportation journals. The goal of this section is not to present a review of TSP and VRP solution methods but to focus on the literature that deals with the estimation of distances in TSPs and VRPs. Comprehensive reviews of solution methods for TSPs and VRPs are given by Gutin and Punnen (*1*) and Toth and Vigo (*2*), respectively.

A seminal contribution to an estimate of the length of a shortest closed path or tour through a set of points was established by Beardwood et al. (*3*). These authors demonstrated that for a set $\boldsymbol{V}^n$ with n points distributed in an area A, the length of the TSP tour through the set $\boldsymbol{V}^n$ asymptotically converges to

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) = k\sqrt{nA} \tag{1}$$

The value of k is a constant. The asymptotic validity of this formula for TSPs was experimentally tested by Ong and Huang (*4*) by using a nearest neighbor and exchange improvement heuristics. With a Euclidean metric and a uniform distribution of customers, the constant term has been estimated at $k = 0.765$ (*5*). For reasonably compact and convex areas, the limit provided by Equation 1 converges

Maseeh College of Engineering and Computer Science, Portland State University, P.O. Box 751, Portland, OR 97207-0751. figliozzi@pdx.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2089, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 1–8.
DOI: 10.3141/2089-01

rapidly (*6*). In compact and convex areas, the following approximation formula can be used:

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) \approx 0.765\sqrt{nA} \tag{2}$$

Equation 2 requires a Euclidean travel metric or L_2 metric. Jaillet (*7*) estimated the constant $k \approx 0.97$ for the Manhattan travel metric or L_1 metric.

Approximations to the length of capacitated VRPs (CVRPs) were first published in the late 1960s and early 1970s (*8–10*). Webb studied the correlation between route distance and customer–depot distances. Eilon et al. (*10*) proposed several approximations to the length of the CVRP based on the shape and area of delivery, the average distance between customers and the depot, the capacity of the vehicle in terms of the number of customers that can be served per vehicle, and the area of a rectangular delivery region.

Daganzo (*11*) proposed a simple and intuitive formula for the CVRP when the depot is not necessarily located in the area that contains the customers:

$$\mathrm{CVRP}\left(\boldsymbol{V}^n\right) \approx 2\bar{r}n/C + 0.57\sqrt{nA} = 2\bar{r}m + 0.57\sqrt{nA} \tag{3}$$

CVRP($\boldsymbol{V}^n$) is the total distance of the CVRP problem serving n customers, the average distance between the customers and the depot is $\bar{r}$, and the maximum number of customers that can be served per vehicle is C. Hence, the number of routes m is known a priori and can be calculated as n/C. Equation 3 can be interpreted as having (*a*) a term related to the distance between the depot and customers and (*b*) a term related to the distance between customers. The coefficients of Equation 3 were derived by assuming $C > 6$ and $N > 4C^2$. Daganzo's approximation works better in elongated areas as the routes were formed following the "strip" strategy. Robusté et al. (*12*) use simulations to analyze elliptical areas and propose adjustments based on area shape, vehicle capacity, and number of customers. Erera's dissertation (*13*) proposes continuous approximations to estimate expected detour and distances in the stochastic version of the CVRP.

Chien (*14*) carried out simulations and linear regressions to test the accuracy of various models to estimate the length of TSP. Chien tested rectangular areas with eight length–width ratios ranging from 1 to 8 and circular sectors with eight central angles ranging from 45 to 360 degrees. Exact solutions to solve the TSPs were used, and the size of the problems is five to 30 customers. The depot was always located at the origin, the left-lower corner of the rectangular areas. Chien randomly generated test problems and by using linear regressions found the best-fitting parameters. The mean absolute percentage error (MAPE) was the benchmark to compare specifications. Chien found that the lowest MAPE for the best model is equal to 6.9%.

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) \approx 2.1\bar{r} + 0.67\sqrt{nR} \qquad R^2 = 0.99 \qquad \mathrm{MAPE} = 6.9 \tag{4}$$

Chien used the area of the smallest rectangle that covers the customers; this area is denoted R. Equation 4 is not convenient for planning purposes when there may be many possible subsets of customers that are not known a priori. The previous models were also estimated for each of the 16 different regions; R^2 and MAPE are reported for each type of region and model. The estimated parameters change according to the shape of the region.

Kwon et al. (*15*) also carried out simulations and linear regressions, but in addition they used neural networks to find better approximations. To test the accuracy of different models, they tested TSPs in rectangular areas with eight length–width ratios ranging from 1 to 8. Models were estimated with the depot located at the origin and at the middle of the rectangle. The sizes of the problems ranged from 10 to 80 customers. Kwon et al. compared Equation 4 with two additional approximations that make use of the geometric information proportioned by the length–width ratio (S) of the rectangle (length and width defined in such a way that the ratio is always larger than or equal to 1). The results obtained for the depot located at the origin are as follows:

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) \approx \left[0.83 - 0.0011(n+1) + 1.11S/(n+1)\right]\sqrt{nA}$$
$$R^2 = 0.99 \qquad \mathrm{MAPE} = 3.71 \tag{5}$$

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) \approx 0.41\bar{r} + \left[0.77 - 0.0008(n+1) + 0.90S/(n+1)\right]\sqrt{nA}$$
$$R^2 = 0.99 \qquad \mathrm{MAPE} = 3.61 \tag{6}$$

Accounting for the shape of the area improves accuracy, although this is at the expense of adding one and two extra terms in the last two expressions. R^1 is defined as the area of the smallest rectangle that covers the customer and the depot. With the depot located at the center of the rectangle, the results obtained are as follows:

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) \approx \left[0.87 - 0.0016(n+1) + 1.34S/(n+1)\right]\sqrt{nR^1}$$
$$R^2 = 0.99 \qquad \mathrm{MAPE} = 3.88 \tag{7}$$

$$\mathrm{TSP}\left(\boldsymbol{V}^n\right) \approx 1.15\bar{r} + \left[0.79 - 0.0012(n+1) + 0.97S/(n+1)\right]\sqrt{nR^1}$$
$$R^2 = 0.99 \qquad \mathrm{MAPE} = 3.70 \tag{8}$$

MAPE slightly increases when the depot is located at the center of the rectangle. Kwon et al. (*15*) also used neural networks to find a model that better predicts TSP length. They concluded that the capability of neural networks to find "hidden" relationships provides a slight edge against regression models. However, the models are less parsimonious and the terms harder to interpret geometrically.

APPROXIMATIONS PROPOSED AND TESTED

In the preceding section, approximations and simulation results for the TSP were reviewed. There are strong theoretical and intuitive reasons to include both $\sqrt{nA}$ and $\bar{r}$ terms in the models. Although Daganzo (*11*) and Robusté et al. (*12*) propose distance formulas for the CVRP, the number of necessary routes or vehicles, m, is known a priori. Daganzo and Robusté et al. assume demands that are factors of the vehicles' capacities; with this assumption the number of vehicles can be accurately determined by using $m = n/C$. In this research it is also assumed that the number of routes needed is known a priori.

Given n customers and m routes, there is a relationship between the number of links that connect the depot and the first or last customer of each route and the number of local intercustomer links. Any solution to a TSP with n customers uses $n + 1$ links, where $n - 1$ links are local and two links are connecting. If capacity or window constraints are added, the resulting VRP has $m \geq 1$ routes. In general, for m routes and n customers, any solution to a VRP uses $n + m$ links. In general, for any given n and m, the number of connecting links is $2m$ and the number of local links is $n - m$.

Six approximations or models to estimate the length of VRP instances are proposed:

$$\text{Model 1: VRP}(\boldsymbol{V}) \approx k_l\sqrt{An} + 2\bar{r}m \quad (9)$$

$$\text{Model 2: VRP}(\boldsymbol{V}) \approx k_l\frac{n-m}{n}\sqrt{An} + 2\bar{r}m \quad (10)$$

$$\text{Model 3: VRP}(\boldsymbol{V}) \approx k_l\sqrt{An} + k_m m \quad (11)$$

$$\text{Model 4: VRP}(\boldsymbol{V}) \approx k_l\frac{n-m}{n}\sqrt{An} + k_m m \quad (12)$$

$$\text{Model 5: VRP}(\boldsymbol{V}) \approx k_l\sqrt{An} + k_b\sqrt{A/n} + k_m m \quad (13)$$

$$\text{Model 6: VRP}(\boldsymbol{V}) \approx k_l\frac{n-m}{n}\sqrt{An} + k_b\sqrt{A/n} + k_m m \quad (14)$$

The parameters k_l, k_b, and k_m are estimated by linear regression. The term $(n - m)/n$ is proposed in this research to modify the local tour distance in Models 2, 4, and 6. This term has some desirable properties: (*a*) when $n = m$ the estimated local distance is zero, whereas (*b*) when $n >> m$ or $m = 1$ the local tour distance tends to the expression suggested by Beardwood et al. (*3*). This research proposes the use of these ideas to reflect the trade-offs between n and m and improve the accuracy of the average VRP distance estimation as a function of n and m.

Increased accuracy can be obtained if additional terms related to the shape of the region and customers are added, as do Kwon et al. (*15*), or if a third term is added, as in Equations 5 and 6:

$$\sqrt{nA}/(n+1) \approx \sqrt{A/n}$$

The functional form of this term was not justified by Kwon et al. (*15*). This research conjectures that this term is a proxy for the average distance from the depot to the first customer plus the distance from the last customer to the depot. It has been shown (*16*) that the average distance to the nearest neighbor is a function of $1/\sqrt{n}$. The term $k_b\sqrt{A/n}$ may be significant in estimating distances for the TSP when n is small and $m = 1$.

The term $k_m m$ estimates the connecting distance and captures increases in connecting distance as m increases or as the depot moves away from the customers. The proposed models, Expressions 9 through 14, are evaluated by using numerical experiments in the section on analysis and discussion of experimental results. The next section describes the experimental setting.

EXPERIMENTAL SETTING

This research utilizes the classical instances of the VRP with time windows proposed by Solomon (*17*) to test the models. The Solomon instances include distinct spatial customer distributions, vehicles' capacities, customer demands, and customer time windows. These problems have been widely studied in the operations research literature, and in addition the data sets are readily available. (Several websites maintain downloadable data sets of the instances, including Solomon's own website: web.cba.neu.edu/~msolomon/problems.htm.)

In the Solomon problems there are 100 customers per instance. The distances and travel times are Euclidean. There are six classes of problems, depending on the geographic location of customers (R, random; C, clustered; RC, mixed random and clustered) and time windows length (1, short time windows; 2, long time windows). The customer coordinates are identical for all problems within one type (i.e., R, C, and RC). The sets R1, C1, and RC1 have vehicle capacity of 200 units, allowing fewer customers per route than the remaining sets. In contrast, Problem Sets R2, C2, and RC2 have vehicle capacity equal to 1,000, 700, and 1,000 units, respectively, allowing a larger number of customers per route. Because of the short time windows, Problem Sets R1, C1, and RC1 allow only a few customers per route (5 to 10). Problem sets R2, C2, and RC2 have longer time windows, and route sizes are on the order of 30 customers per route. The first instance of each problem class is used in this research.

Random samples of the Solomon problems are used to examine the accuracy of Models 9 through 14. Out of $N = 100$ possible customers in a service area A, a problem or instance is formed by a subset of n randomly selected customers. On the basis of the first instance of the six problem types proposed by Solomon, 15 subsets of customers of size 70, 60, 50, 40, 30, 20, and 10 were randomly selected from the original 100 customers. To incorporate different levels of customer demand, new instances were created by applying the demand factors presented in Table 1 to each subset of customers. On the basis of the factors in the second row of demand factors in Table 1, the customers have demands similar to those in the original Solomon problems. The resulting problems using the highest demand multipliers (last row of Table 1) are such that some customers are truckload (TL) or almost TL customers. Increasing some customer demands to or close to the TL level was done to test the models when problems are highly constrained and have a large number of routes and a small number of customers per route. On the other hand, the situation of having a large number of customers per route is obtained when the demand factor is zero (first row Table 1). In all cases the routes' durations were limited by the depot time window. Finally, in all Solomon problems, customers' time windows are different in width and start time. This adds a layer of variability. Hence, for each problem class or set, variability is introduced in three distinct ways: different

TABLE 1 Truck Capacity and Customer Demand Data by Problem Type

	Instance					
	C1	R1	CR1	C2	R2	CR2
Vehicle capacity	200	200	200	700	1,000	1,000
Max. demand	50	41	40	41	41	40
Demand factors	0	0	0	0	0	0
	1	1	1	1	1	1
	1.6	1.78	1.8	3.6	5.68	5.8
	2.2	2.56	2.6	6.2	10.36	10.6
	2.8	3.34	3.4	8.8	15.04	15.4
	3.4	4.12	4.2	11.4	19.72	20.2
	4	4.9	5	14	24.4	25

subsets of customer locations, different levels of customer demands, and nonuniform time windows.

Most studies have focused on the derivation or testing of asymptotic estimators of the length TSPs (*3, 4*). Hence, experimental tests have mostly included a large number of customers per route. However, real-life routes have a relatively small number of customers per route because of capacity or tour length constraints. For example, in Denver, Colorado, more than 50% of single and combination truck routes include less than six stops (*18*); 95% of the truck routes include less than 20 stops. This research work tests the models by using instances that range from one customer per route to more than 35 customers per route.

In the Solomon problems the depot has a central location with respect to the customers. To test the model when the depot is located in the periphery, all the created instances were also solved with the depot located at the origin [i.e., coordinates (0, 0)]. To study the model quality and parameter values without time windows, all the problem instances were also solved without time windows. To the best of the author's knowledge, there is no published research reporting MAPE and simulation results for CVRP or VRP with time windows.

All problem instances in this research were solved with a VRP improvements heuristic that has obtained the best published solution in terms of number of vehicles (*19*). The solution quality of this heuristic is clearly superior to the performance of savings or construction heuristics used in previous research efforts such as those of Ong and Huang (*4*) or Robusté et al. (*12*). The overall solution quality of the heuristic used—that is, the total number of vehicles needed to solve the 56 Solomon problems—is approximately 4% over the best known solutions (*20, 21*).

To evaluate the prediction accuracy, the MAPE and the mean percentage error (MPE) are used. They are calculated as follows:

$$\text{MPE} = \frac{1}{p}\sum_{i=1}^{p}\frac{D_i - E_i}{D_i}\times 100\%$$

$$\text{MAPE} = \frac{1}{p}\sum_{i=1}^{p}\frac{\left|D_i - E_i\right|}{D_i}\times 100\%$$

where the actual distance for instance i is denoted D_i and the estimated distance is denoted E_i.

For a given set of instances it is always the case that MPE ≤ MAPE. The MPE indicates whether the estimation, on average, overestimates or underestimates the actual distance. The MAPE provides the average deviation between actual and estimated distance as a percentage of the actual distance.

ANALYSIS AND DISCUSSION OF EXPERIMENTAL RESULTS

Results for CVRP instances (i.e., no time windows) and the depot located at the center are shown in Table 2. All the regression results were obtained by forcing the intercept or constant term to be zero; this is consistent with previous studies by Chien (*14*) and Kwon et al. (*15*). In the regression models, the average distance per sample size is the dependent variable. Model fit R^2, MAPE, and MPE are displayed for Models 1 through 6. The average, maximum, and minimum correspond to the first Solomon problem in each of the six problem types (R1, C1, RC1, R2, C2, and RC2). For the sake of clarity, only three decimals are displayed.

TABLE 2 Model Fit Comparison with a Central Depot and No Time Windows

Model	Statistic	R^2	MPE (%)	MAPE (%)
Model 1	Average	0.966	1.4	6.0
	Min	0.933	–0.8	4.2
	Max	0.986	3.5	7.3
Model 2	Average	0.991	1.5	4.7
	Min	0.986	–1.2	3.1
	Max	0.994	4.2	6.5
Model 3	Average	0.999	1.0	4.0
	Min	0.998	–0.9	2.2
	Max	1.000	3.5	6.4
Model 4	Average	0.999	–0.7	3.2
	Min	0.999	–2.6	1.7
	Max	1.000	1.6	4.5
Model 5	Average	0.999	–0.4	3.1
	Min	0.999	–0.7	2.0
	Max	1.000	–0.1	4.3
Model 6	Average	1.000	–0.1	2.4
	Min	0.999	–0.3	1.5
	Max	1.000	0.1	3.4

In Table 2, all six models have good R^2 values. However, models with more terms (such as Models 5 and 6) have a superior MAPE performance. The models that adjust the tour distances by using the term $(n - m)/n$ (Models 2, 4, and 6) have a better MAPE performance than do their counterparts with the same number of estimated coefficients (Models 1, 3, and 5, respectively).

Table 3 indicates the impact of time windows on the accuracy of average distance estimation. These results were obtained by using the same instances used previously to obtain Table 2 but considering all the customer time windows as originally intended in the Solomon problems. A slight decrease in the R^2 values is observed. Imposing time windows decreases the predictive ability of all six models. The increases in MAPE range from 50% to 175% for Model 3. As observed in Table 2, the models that adjust the tour distances by using the term $(n - m)/n$ have a better MAPE performance than do their counterparts.

TABLE 3 Model Fit Comparison with a Central Depot and Time Windows

Model	Statistic	R^2	MPE (%)	MAPE (%)
Model 1	Average	0.968	6.2	12.0
	Min	0.954	3.2	6.2
	Max	0.982	10.2	17.9
Model 2	Average	0.984	4.9	7.9
	Min	0.977	2.5	5.2
	Max	0.990	9.2	12.7
Model 3	Average	0.994	6.8	11.0
	Min	0.987	3.5	5.8
	Max	0.998	12.3	17.8
Model 4	Average	0.997	4.3	6.9
	Min	0.994	1.6	2.8
	Max	0.999	8.8	12.1
Model 5	Average	0.998	–0.3	4.8
	Min	0.997	–0.8	2.8
	Max	0.999	0.2	7.1
Model 6	Average	0.999	–0.1	3.7
	Min	0.998	–0.5	2.1
	Max	1.000	0.3	5.7

TABLE 4 Local Tour Regression Coefficient Without Time Windows (Model 2)

Instance	Coeff.	*t*-Stat.	St. Error
C101	0.62	41.58	0.01
R101	0.87	37.40	0.02
RC101	0.79	25.82	0.03
C201	0.64	57.71	0.01
R201	0.90	48.85	0.02
RC201	0.80	33.79	0.02
Average	0.77	40.86	0.02

Furthermore, the performance of Models 2, 4, and 6 is better with time windows compared with that of Models 1, 3, and 5. This can be explained by the larger number of routes needed when time windows are introduced; when m is larger, the term $(n - m)/n$ plays a more significant role.

Time windows also affect the value of the estimated local tour parameter k_l. Table 4 shows the value of the parameter k_l for Model 2 and customers without time window constraints. The value of k_l changes with the spatial distribution of the customers; it is highest for randomly distributed instances and lowest for clustered instances. This is intuitively correct since the value of k_l is a proxy for the average distance between customers in a local tour (between the first and last customer of a route). Table 5 shows the value of the parameter k_l for Model 2 and customers with time window constraints. In comparison with the values in Table 4, all parameters k_l show an increase that is highly statistically significant. This is intuitively correct since time window constraints do not allow the formation of compact routes; hence, the average distance between customers in the local tours almost doubles.

The same models were also estimated with the depot located at the corner [i.e., coordinates (0, 0)]. Moving the depot to the corner increases the average distance between the depot and the customers considerably. Tables 6 and 7 show the results with and without time windows and the depot at a corner. Despite the change in the depot location, the same trends are still observed: (*a*) the models adjusted by $(n - m)/n$ perform better in terms of MAPE than their counterparts, (*b*) time windows decrease the predictive accuracy of the models, and (*c*) with time windows the parameter k_l increases. With the corner depot, all three models perform better in terms of MPE and MAPE than with a centrally located depot. The same phenomenon can be observed in the experimental results of Kwon et al. (*15*) for TSP distances.

TABLE 5 Local Tour Regression Coefficient with Time Windows (Model 2)

Instance	Coeff.	*t*-Stat.	St. Error
C101	1.30	56.24	0.02
R101	1.39	45.80	0.03
RC101	1.06	45.17	0.02
C201	1.32	69.79	0.02
R201	1.89	53.47	0.04
RC201	1.74	63.38	0.03
Average	1.45	55.64	0.03

TABLE 6 Model Fit Comparison with a Corner Depot and No Time Windows

Model	Statistic	R^2	MPE (%)	MAPE (%)
Model 1	Average	0.985	0.8	3.3
	Min.	0.970	−1.0	1.7
	Max.	0.994	2.8	4.5
Model 2	Average	0.981	1.4	2.9
	Min.	0.965	−0.2	1.5
	Max.	0.995	3.0	4.7
Model 3	Average	1.000	−1.0	3.1
	Min.	0.999	−2.3	1.9
	Max.	1.000	0.9	4.4
Model 4	Average	1.000	0.1	2.1
	Min.	0.999	−1.0	1.7
	Max.	1.000	1.2	2.6
Model 5	Average	1.000	−0.4	2.1
	Min.	0.999	−0.7	1.3
	Max.	1.000	−0.1	2.6
Model 6	Average	1.000	−0.1	1.7
	Min.	0.999	−0.3	1.2
	Max.	1.000	0.0	2.2

The value of k_m is closer to the corresponding value of $2\bar{r}$ when the depot is not centrally located. With a central depot, the value of the parameter k_m in Models 3 through 6 is within approximately 20% of the corresponding value of $2\bar{r}$; with a corner depot, the value of the parameter k_m in Models 3 through 6 is within approximately 10% of the corresponding value of $2\bar{r}$. Hence, the average distance from the depot to the customers, $2\bar{r}$, still appears in Models 3 through 6 but under the form of the estimated coefficient k_m.

The comparisons of the results in Tables 2, 3, 6, and 7 indicate that Model 6 is clearly superior in terms of MPE and MAPE across all experimental settings. However, the improved accuracy requires the estimation of a larger number of parameters (three). In addition, the interpretation of the term $k_b\sqrt{A/n}$ is not straightforward. The sign of the estimated k_b parameter is positive in all instances with a centrally located depot, but it is negative in some instances with a

TABLE 7 Model Fit Comparison with a Corner Depot and Time Windows

Model	Statistic	R^2	MPE (%)	MAPE (%)
Model 1	Average	0.955	3.4	8.7
	Min.	0.941	0.0	5.6
	Max.	0.977	6.2	13.2
Model 2	Average	0.985	3.2	5.4
	Min.	0.976	0.9	3.2
	Max.	0.989	6.6	9.3
Model 3	Average	0.998	2.7	5.6
	Min.	0.996	1.0	2.4
	Max.	0.999	6.2	10.2
Model 4	Average	0.999	3.0	5.0
	Min.	0.997	1.0	3.5
	Max.	0.999	6.1	9.0
Model 5	Average	0.999	−0.4	3.9
	Min.	0.999	−0.9	2.1
	Max.	1.000	−0.1	5.6
Model 6	Average	0.999	−0.2	2.9
	Min.	0.999	−0.5	1.6
	Max.	1.000	0.0	4.5

corner depot. The sign change takes place in both Models 5 and 6. A plausible explanation is that the term $k_b\sqrt{A/n}$ captures part of the average distance to or from the depot when the depot is located inside the customer service area; this is congruent with the assumptions of the analytical expression provided by Clark and Evans (*16*) for the average nearest neighbor distance.

Model 2 is superior if parsimony and interpretability in addition to accuracy are taken into account. Model 2 is simple and easily interpreted as well as robust when time windows are introduced. The coefficient k_b is in all cases easily interpreted, highly significant, and positive. Furthermore, Model 2 outperforms Model 3, which uses two regression coefficients with time windows and a corner depot. This can be attributed to the influence of the term $(n - m)/n$, which plays a larger role when more routes are required (more routes are required when time windows are introduced and when the depot is moved away from the customers).

REAL-LIFE APPLICATION

Previous literature has only tested TSP or CVRP distance approximations on simulated environments with Euclidean distances. Although approximation formulas have theoretical applications in transport and logistics planning models, they can also be used to estimate distance, costs, and times in real-life planning applications. The original motivation for this research came from the study of distribution routes for a freight forwarding company based in Sydney, Australia. Distribution tours originated at a depot located close to the Port of Sydney; the customers were mostly located in different industrial suburbs. The pattern of customer distribution resembles the mix of random and clustered customers as in the random–clustered Solomon problems. The company's customers are in the hundreds, but they are not visited every day. The freight forwarding company consolidates less-than-container shipments, and customers are visited only if a consignment has arrived before the distribution cutoff time. Further details about the tour characteristics are given by Figliozzi (*22*).

Model 4 was tested with customers located in the industrial suburb of Bankstown with 30 customers distributed in an irregular area of 39.5 km^2 (see Figure 1). The delivery area is bordered by the Bankstown local airport in the west, a freeway in the south, and secondary highways in the east and north. The average distance between the depot and the industrial suburb is approximately 22 km on the connecting freeway. To test Model 4, five sets of 2, 4, 6, 8, 10, 15, and 20 customers were randomly chosen among the existing customers in the suburb to simulate the daily demand. Selecting random subsets of customers from the pool of existing customers in the area is a fair representation of the real demand. The number of customers visited per day varies widely; it may be as low as one or two or, exceptionally, close to 30. In the results presented below, all customers have the same probability of a visit. Although this is not the case in reality, it simplifies the exposition and introduces greater variation in the customer subsets.

Because of contract and labor policies, the main distribution cost is associated with the number of driver hours needed. Therefore, the objective is to minimize total route durations, avoiding expensive overtime (the overtime pay rate is 50% higher). An important consideration in working with travel times in an urban area is that speeds are strongly influenced by congestion, road characteristics, and speed limits. In this application the travel speeds used are 65 km/h on freeways, 35 km/h on main connecting streets (four lanes or more with traffic lights), and 25 km/h on local streets. With this speed information, a matrix of shortest travel times between customers and depot was constructed by using the urban highway network and geographic information system software. Figure 2 shows the relationship

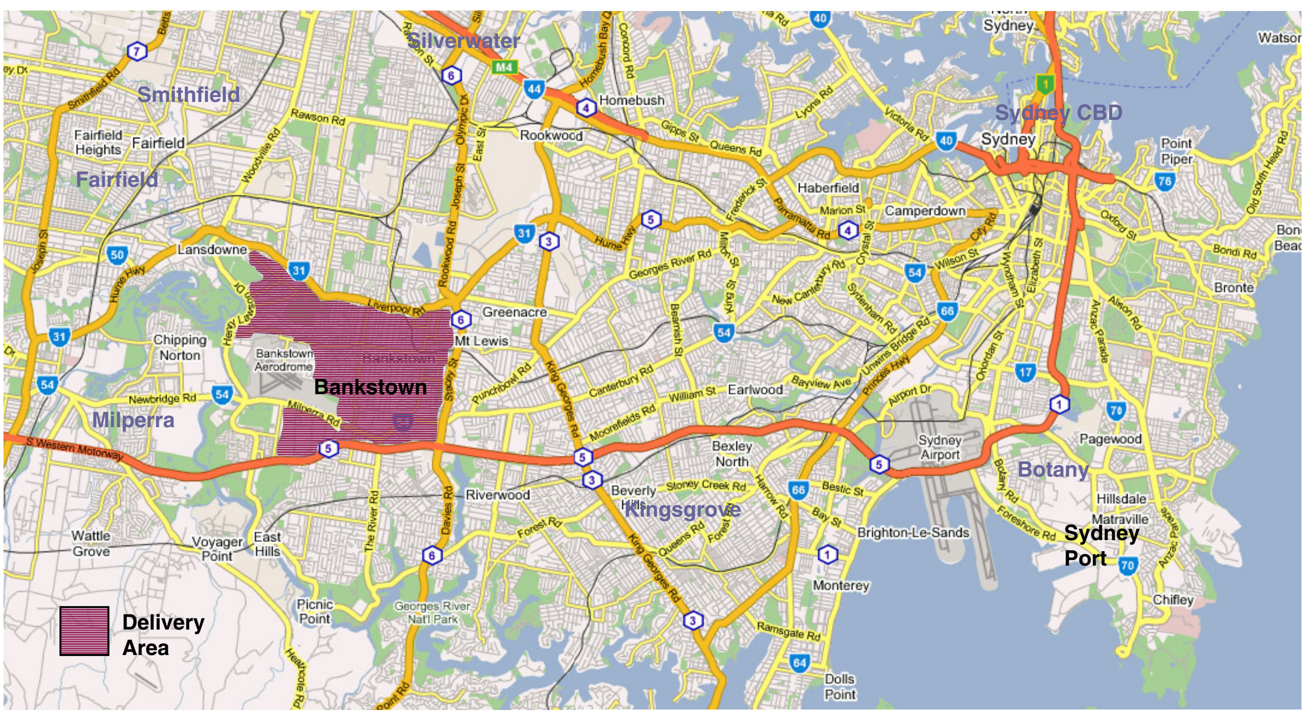

FIGURE 1 Relative location of the Port of Sydney and delivery industrial areas.

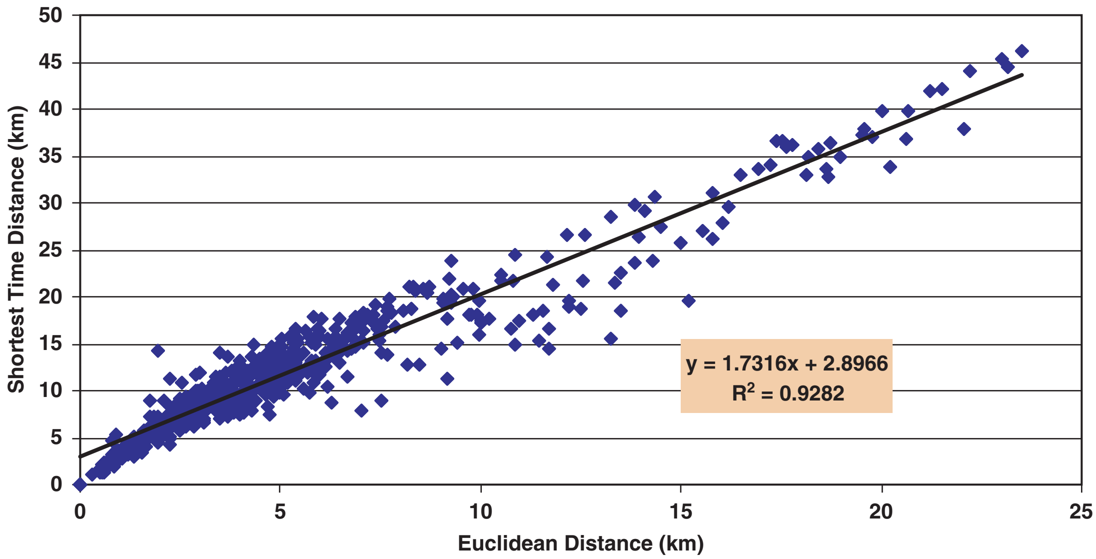

FIGURE 2 Euclidean distance versus shortest time distance among customers and depot customers.

between the Euclidean distance and the distance based on the shortest time path for all customers and the depot. The high concentration of short distance points close to the origin corresponds to the distances between customers in the suburb, while the longer distances are mostly depot–customer. The R^2 of 0.93 indicates that despite the irregular shape of the distribution area and the mix of travel speeds, the Euclidean distance is a fairly good predictor of the actual distance traveled between customer pairs or customer–depot pairs. From existing customer data, an average service time of 45 min per customer is used.

Three routing scenarios were constructed: (*a*) no constraints or TSP case, (*b*) with a tour duration constraint of 8 h, and (*c*) adding 4-h time windows per customer. The number of routes varied from one in the TSP instances to five in the instances with time windows. The regression was estimated with the consolidated data from all three scenarios. The results are shown in Table 8. The network distance traveled is well approximated with a MAPE of 4.2%. The prediction of travel time or driving time in hours has a MAPE of 11.7%. The good MAPE is not surprising given the good correlation between distance traveled and time driven (see Figure 3). Model 4 was used to approximate times and distances due to the different travel speeds; the connecting distance between depot and customers does not always follow the same type of highway. These results are encouraging and show that the proposed models may have useful applications in urban networks and modeling applications (*23, 24*). While these results are promising, from this example it is impossible to generalize the results. Further research is necessary to study the accuracy of VRP distance approximation in cities with different layouts and highway networks.

CONCLUSIONS

This research studies approximations to the average length of VRPs when there is variability in the number, level, and locations of customer demands. The approximations are intended for strategic and planning analysis of transportation and logistics problems, when the number and location of customers vary daily and are not known a priori.

A new parsimonious, intuitive, and effective approximation is proposed and successfully tested by using instances with different patterns of customer spatial distribution, time windows, customer demands, and depot locations. It was found that time windows negatively affect the accuracy of the approximations. Time windows increase travel distance not only because the number of routes is increased but also because the separation between customers per route is increased. As the distance between the depot and delivery region increases, the accuracy of the approximation increases. The approximation was also tested in a real-life urban network with encouraging results.

TABLE 8 Real-Life Network Distance and Time Estimation (Model 4)

					Coefficient	
Instance	R^2	MPE (%)	MAPE (%)		k_l	k_m
Distance	0.999	–0.5	4.2	Estimated	0.80 km	49.51 km
				t-stat.	4.158	48.317
Time driven	0.988	5.9	11.7	Estimated	0.028 h	1.25 h
				t-stat.	2.838	13.088

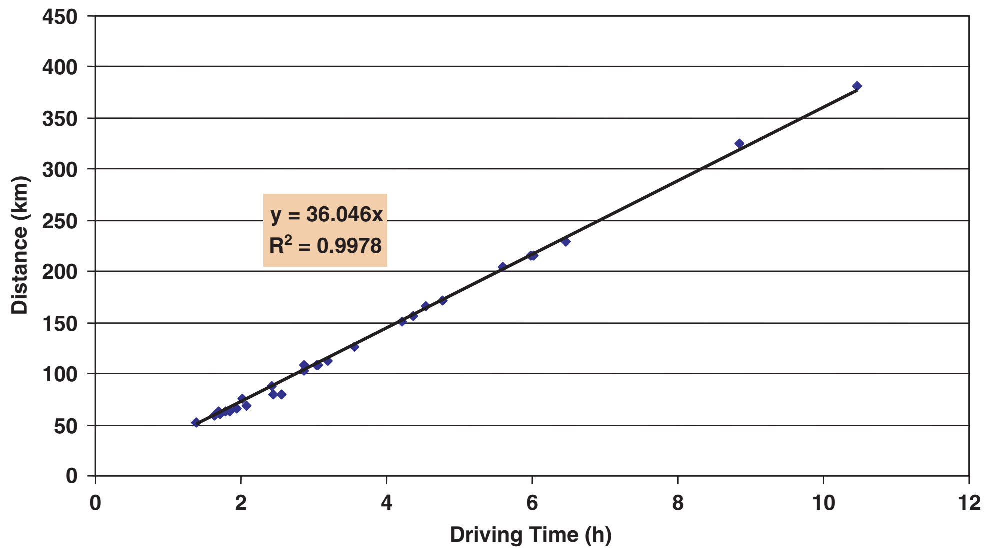

FIGURE 3 Distance traveled and time driven.

REFERENCES

1. Gutin, G., and A. Punnen (eds.). *The Traveling Salesman Problem and Its Variations.* Kluwer, Boston, Mass., 2002.
2. Toth, P., and D. Vigo (eds.). *The Vehicle Routing Problem.* Society for Industrial and Applied Mathematics, 2001.
3. Beardwood, J., J. H. Halton, and J. M. Hammersley. The Shortest Path Through Many Points. *Proceedings of the Cambridge Philosophical Society,* Vol. 55, 1959, pp. 299–327.
4. Ong, H. L., and H. C. Huang. Asymptotic Expected Performance of Some TSP Heuristics: An Empirical Evaluation. *European Journal of Operational Research,* Vol. 43, No. 2, 1989, pp. 231–238.
5. Stein, D. An Asymptotic Probabilistic Analysis of a Routing Problem. *Mathematics of Operations Research,* Vol. 3, No. 2, 1978, pp. 89–101.
6. Larson, R. C., and A. R. Odoni. *Urban Operations Research.* Prentice-Hall, Inc., 1981.
7. Jaillet, P. A Priori Solution of a Traveling Salesman Problem in Which a Random Subset of the Customers Are Visited. *Operations Research,* Vol. 36, No. 6, 1988, pp. 929–936.
8. Webb, M. Cost Functions in the Location of Depots for Multiple Delivery Journeys. *Operational Research Quarterly,* Vol. 19, 1968, pp. 311–315.
9. Christofides, N., and S. Eilon. Expected Distances in Distribution Problems. *Operational Research Quarterly,* Vol. 20, No. 4, 1969, pp. 437–443.
10. Eilon, S., D. Watson-Gandy, and N. Christofides. *Distribution Management: Mathematical Modelling and Practical Analysis.* Hafner, New York, 1971.
11. Daganzo, C. F. The Distance Traveled to Visit *N* Points with a Maximum of *C* Stops per Vehicle: An Analytic Model and an Application. *Transportation Science,* Vol. 18, No. 4, 1984, pp. 331–350.
12. Robusté, F., M. Estrada, and A. López-Pita. Formulas for Estimating Average Distance Traveled in Vehicle Routing Problems in Elliptic Zones. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1873,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 64–69.
13. Erera, A. *Design of Large-Scale Logistics Systems for Uncertain Environments.* PhD dissertation. University of California, Berkeley, 2000.
14. Chien, T. W. Operational Estimators for the Length of a Traveling Salesman Tour. *Computers and Operations Research,* Vol. 19, No. 6, 1992, pp. 469–478.
15. Kwon, O., B. Golden, and E. Wasil. Estimating the Length of the Optimal TSP Tour: An Empirical Study Using Regression and Neural Networks. *Computers and Operations Research,* Vol. 22, No. 10, 1995, pp. 1039–1046.
16. Clark, P. J., and F. C. Evans. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology,* Vol. 35, No. 4, 1954, pp. 445–453.
17. Solomon, M. M. Algorithms for the Vehicle-Routing and Scheduling Problems with Time Window Constraints. *Operations Research,* Vol. 35, No. 2, 1987, pp. 254–265.
18. Holguín-Veras, J., and G. R. Patil. Observed Trip Chain Behavior of Commercial Vehicles. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1906,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 74–80.
19. Figliozzi, M. An Iterative Route Construction and Improvement Algorithm for the Vehicle Routing Problem with Soft and Hard Time Windows. *Proc., Applications of Advanced Technologies in Transportation Conference,* Athens, Greece, May 2008.
20. Braysy, I., and M. Gendreau. Vehicle Routing Problem with Time Windows, Part I: Route Construction and Local Search Algorithms. *Transportation Science,* Vol. 39, No. 1, 2005, pp. 104–118.
21. Braysy, I., and M. Gendreau. Vehicle Routing Problem with Time Windows, Part II: Metaheuristics. *Transportation Science,* Vol. 39, No. 1, 2005, pp. 119–139.
22. Figliozzi, M. A. Commercial Vehicle Tours and Road Congestion in Urban Areas: Implications for Carriers' Operations and Public Data Collection and Dissemination Efforts. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
23. Figliozzi, M. A. Analysis of the Efficiency of Urban Commercial Vehicle Tours: Data Collection, Methodology, and Policy Implications. *Transportation Research B,* Vol. 41, No. 9, 2007, pp. 1014–1032.
24. Figliozzi, M. A. The Impacts of Congestion on Commercial Vehicle Tours Characteristics and Costs. *Proc., 2nd Annual National Urban Freight Conference,* Long Beach, Calif., Dec. 2007.

The Transportation Network Modeling Committee sponsored publication of this paper.