Title
Pickup and Delivery Systems For Overnight Carriers

Author
Hall, Randolph W.

Publication Date
1992-07-01

Pickup and Delivery Systems For Overnight Carriers

Randolph W. Hall

July 1992
Working Paper, No. 106

The University of California
Transportation Center

University of California
Berkeley, CA 94720

The University of California
Transportation Center

The University of California Transportation Center (UCTC) is one of ten regional units mandated by Congress and established in Fall 1988 to support research, education, and training in surface transportation. The UC Center serves federal Region IX and is supported by matching grants from the U.S. Department of Transportation, the California State Department of Transportation (Caltrans), and the University.

Based on the Berkeley Campus, UCTC draws upon existing capabilities and resources of the Institutes of Transportation Studies at Berkeley, Davis, and Irvine; the Institute of Urban and Regional Development at Berkeley; the Graduate School of Architecture and Urban Planning at Los Angeles; and several academic departments at the Berkeley, Davis, Irvine, and Los Angeles campuses Faculty and students on other University of California campuses may participate in Center activities. Researchers at other universities within the region also have opportunities to collaborate on selected studies. Currently faculty at California State University, Long Beach, and at Arizona State University, Tempe, are active participants.

UCTC's educational and research programs are focused on strategic planning for improving metropolitan accessibility, with emphasis on the special conditions in Region IX. Particular attention is directed to strategies for using transportation as an instrument of economic development, while also accommodating to the region's persistent expansion and while maintaining and enhancing the quality of life there.

The Center distributes reports on its research in working papers, monographs, and in reprints of published articles. For a list of publications in print, write to the address below.

University of California
Transportation Center

108 Naval Architecture Building
Berkeley, California 94720
Tel. 415/643-7378
FAX· 415/643-5456

Authors of papers reporting on UCTC-sponsored research are solely responsible for their content This research was supported by the U.S Department of Transportation and the California State Department of Transportation, neither of which assumes liability for its content or use

Pickup and Delivery Systems For Overnight Carriers

Randolph W. Hall

Department of Industrial Engineering and Operations Research
Institute of Transportation Studies
University of California at Berkeley

Working Paper, No. 106

The University of California Transportation Center
University of California at Berkeley

PICKUP AND DELIVERY SYSTEMS FOR OVERNIGHT CARRIERS

Randolph W. Hall
Department of Industrial Engineering & Operations Research, and
The Institute of Transportation Studies
University of California, Berkeley, CA 94720

ABSTRACT

This paper demonstrates how the constraints of overnight delivery affect the design of pickup and delivery systems. Cost depends on the number of vehicle routes needed to pickup and deliver shipments. This number in turn depends on the critical parts of the driver's day: the morning delivery period, up to the delivery deadline, and the afternoon pickup period, after the cutoff time. Pickup routing is the more complicated, due to the dynamic nature of customer calls. Models are developed to assess the workload remaining to be completed at the cutoff time and, from this value, the number of routes required to pick up the shipments.

ACKNOWLEDGMENTS

My appreciation goes to Chuck Wong of Federal Express, who was helpful in formulating the paper's model.

INTRODUCTION

Transportation deregulation during the 1980s stimulated many American carriers to develop new and innovative ways to move freight. Perhaps the most dramatic success story is Federal Express. A novel airline network structure, integrated with a comprehensive pickup and delivery service, allowed Federal Express to create a new market for overnight mail delivery. Today, Federal Express continues to be the dominant carrier in the American market, though it has been joined by a variety of competitors, ranging from the United Parcel Service to the U.S. Post Office.

Not surprisingly, the Federal Express success story has invited study by academicians. Chan and Ponder (1979), Chestler (1985), Emery et al (1986) and Finnegan and Andrade (1984) reprise the Federal Express experience, focusing on the organizational factors that led to its success. Hall (1989) examined the route structure of overnight carriers, and showed how time zones and time windows affect terminal location and routing strategy. More generally, the phenomenon of the hub-and-spoke network structure has been studied by Kanafani & Ghobrial (1985) and O'Kelly (1986), though the emphasis has been on passenger networks.

To date, there has been no research on the design of pickup and delivery (P&D) systems to support overnight carriers. Without the symbiosis between air and surface transportation, it would be impossible for overnight carriers to provide next day service. Further, P&D comprises a significant percentage of the cost of moving packages, and this percentage is sure to increase as carriers follow Federal Express's recent lead in adopting more decentralized airline network structures. Therefore, design of the P&D system is an

important concern.

The question addressed in this paper is how the constraints of overnight delivery affect the design of the pickup and delivery system. Of particular interest will be the dynamic routing of pickup vehicles (i.e., calls for pickup arrive while the vehicle is in motion). In prior work, Bertsimas and van Ryzin (1990) developed a continuous-space model for the dynamic traveling salesman problem. In their investigation of routing heuristics, they found that the distance traveled per point asymptotically approaches $k/\sqrt{\rho}$, where ρ is the time-average stop density and k is a constant (.64 for a nearest neighbor heuristic with Euclidean metric). The authors also examined queueing effects, measuring the mean time to respond to a call. Jaillet and Odoni (1988) studied a form of dynamic routing where each of a set of stops has a set probability of appearing on a given day. The objective is to construct a fixed route of minimum expected length. Hence, the route is not adapted on any day to reflect actual stop locations. Heuristic algorithms for dynamic routing of cargo ships have been investigated by Psaraftis (1988), who also provide a reasonably complete review of dynamic routing literature.

In overnight P&D, measuring queueing time is less important than satisfying time window constraints. Specifically, when a shipment is picked up in a day has no impact on when it is delivered. The key consideration in P&D design is insuring that a driver's workload does not exceed the amount of time available in his pickup and delivery time windows. At first glance, it may seem that overnight P&D routing is the same as the general problem of routing with time windows (Daganzo, 1987) or the same as dial-a-ride routing (Daganzo, 1978; Stein, 1978). However, unlike the former, stops arrive

dynamically and most customers share a common time window. Unlike the latter, shipments share a common destination (the vehicle terminal) and waiting time to pickup is not the primary concern.

SYSTEM DESCRIPTION

Common carrier networks can be divided into "local area" and "wide area" components. The Wide Area Network (WAN) transports shipments between metropolitan regions. In the case of overnight package carriers, this is usually accomplished by air, via a "one-terminal" topology, i.e., each shipment is only sorted at at one intermediate air terminal between metropolitan regions (see Hall, 1987).

Local Area Networks (LAN) transport shipments within metropolitan regions (Hall, 1991). LANs serve both intra-regional and inter-regional functions, performing pickup and delivery for the former, pickup *or* delivery for the latter. In the case of overnight carriers, the LAN consists of a "Gateway" air terminal -- which connects the WAN to the LAN -- as well as local "stations," where Pickup and Delivery (P&D) vehicles are based.

In the pickup process, overnight carriers serve three types of customers: regular customers, drop boxes, and call-ins. Regular customers act as standing orders, to be visited within a prescribed time window each day. Drop boxes are also served every day, usually at the very end of the route, after a set pickup time.

Customers in the call-in category request pick-ups over the phone through centralized call centers on the day that the pickup is desired. After the

request is received, the operator electronically relays information on the shipment -- address, number of pieces, ready time, closing time, etc. -- to the pickup/delivery station that serves the customer. A dispatcher will determine which driver is assigned to the customer's district, and whether or not the driver is able to handle the call. If the assigned driver is overloaded, the dispatcher has discretion to assign the customer to an alternate. In either case, the dispatcher electronically transmits the shipment information to the selected driver, who will see the call appear on a small CRT in his van. It is up to the driver to select the stop sequence and the routes between the stops.

Overnight carriers operate with location specific cut-off times, typically about 3:30 in the afternoon. To guarantee same-day pickup, customers must call in before the cut-off time. Between the cut-off time and the end of the workday (about 5:00), drivers serve all outstanding call-ins as well as their regular pickups and, finally, the drop boxes. Because many people request pickups immediately prior to the cut-off, and because there are many regular customers, the period from 3:30 to 5:00 is usually the busiest part of the driver's day.

The driver does not return to his/her station until the last pickup of the day is completed. Upon return, shipments are transshipped onto trucks and taken to the gateway terminal. Local shipments (destinations within the metropolitan region) are then separated from inter-regional shipments, and sorted by station.

In the morning, deliveries (both local and those arriving from outside the region) are transported by truck from the airport back to the stations.

In the case of Federal Express, Priority Shipments, which must be delivered by 10:30 a.m., are separated from Regular Shipments, which have later delivery times. From 8:00 until almost 10:30, drivers may work exclusively on priority deliveries, in order to meet their deadlines. This is true even if a regular stop is directly on the driver's delivery route, for even a few minutes delay could cause a priority shipment to be delivered late.

The five-hour period from 10:30 to 3:30 is the lightest of the day. During this time, drivers make regular deliveries as well as early pickups. However, most of the call-ins do not occur until after 1:30 in the afternoon, and few of the regular pick-up stops can be made before 3:30. In some cases, part-time drivers are used, with shifts either ending soon after 10:30 or beginning just before 3:30. However, it can be difficult to hire quality part-time drivers. To make good use of full-timers, it is important that drivers be scheduled to serve both the morning and evening peaks. Counting linehaul and station time, this means that drivers must be on duty from about 7:00 in the morning to 6:00 in the evening. Clearly, this is longer than an ordinary 8-hour day. Federal Express resolves this problem by scheduling drivers for four 10-hour shifts per week, instead of the usual five 8-hour shifts.

Beyond those mentioned already, some important system characteristics follow:

(1) Customers may send and receive multiple pieces. Typically, pickups will contain more pieces than deliveries. Conversely, this means that drivers typically make more delivery stops in a day than pickup stops.

(2) In residential areas, deliveries are much more prevalent than pickups.

(3) In the case of large companies, pickup and delivery sites may be spread throughout a building. Because these sites are usually not coordinated, a driver may have to visit a company several times during a day to make all pickups.

(4) If a distributor is very large, the carrier will sometimes keep a truck parked at the loading dock through the day, or dispatch a special truck to pick up shipments.

(5) The time to process a pickup is much larger than the time to process a delivery, due to time needed to fill out forms.

(6) Delivery routes can be planned in advance, whereas pickup routes must be constantly modified as calls are received.

DESIGN ISSUES

There are many elements to the LAN design, including selecting: (1) the number and location of stations, (2) the topology for transporting shipments between stations, (3) the geometry of pickup and delivery routes, and (4) cutoff and delivery times. The first two issues are discussed in Hall (1991), and will not be covered in this paper. The emphasis here is on the pickup and delivery routes, along with the cutoff and delivery times.

A clear tradeoff exists between the cost of operating the P&D fleet, the quality of service provided and the number of customers attracted based on the service quality. If cutoff times were shifted from 3:30 to 2:00, for instance, drivers would have a longer time interval to serve the end-of-day shipments, and would not have to duplicate routes covered earlier in the day, before all of the calls are received. However, this cost saving would not be appealing if a large portion of the customers shifted to a competitor who offers more responsive service.

The approach taken in this paper is to model the cost for the two critical parts of the day -- the morning delivery period, up until the

delivery deadline, and the afternoon pickup period, surrounding the cutoff. *The paper will assume that pickups are not mixed with deliveries during these periods, and that only priority shipments are delivered before the delivery deadline.* Further, the paper will assume that large customers, that merit dedicated routes, are not incorporated in the pickup/delivery system. Finally, the paper will assume that drop boxes, because they are served last, can be scheduled independently of other pickups.

The major system cost components include: (1) Driver wages and benefits, (2) Vehicle ownership and insurance, (3) Vehicle maintenance, and (4) Fuel. Given a fixed daily shift length, the first cost is a linear function of the number of vehicle routes, whereas the fourth cost is a linear function of the number of miles traveled. The second and third costs depend on a combination of number of vehicle routes and miles traveled. Recognizing that overnight carriers use ordinary vans (not expensive trucks) for P&D, wages and benefits are the dominant cost, and the number of routes is the key factor that determines total P&D cost. Therefore, the question of how to design the P&D system can be phrased in the following way:

How can the number of routes be minimized while meeting service constraints? and
How does the number of routes depend on changes in the service constraint?

MORNING WORKLOAD

This section develops an approximate model to estimate the number of vehicle routes required to deliver priority shipments. The model will assume

that delivery routes meet the objective of minimizing miles traveled, and that only priority shipments are delivered before the priority deadline. In doing so, the number of delivery routes will be minimized. The following symbols will be used:

W_d = size of time window available for deliveries

N_d = minimum number of routes needed for deliveries

$\bar{\rho}_d$ = spatial mean density for delivery stops

$C_{\rho d}$ = coefficient of variation of the spatial mean density, for delivery stops.

s_d = mean time to make a delivery stop (not counting mileage)

v_d = mean vehicle velocity during deliveries

B = area of entire service region among all routes.

The average time required to make a single delivery stop can be expressed as:

$$t_d = s_d + d(\bar{\rho}_d, C_{\rho d})/v_d, \qquad (1)$$

where:

$d(\bar{\rho}_d, C_{\rho d})$ = mean distance between route stops, given $\bar{\rho}_d$ and $C_{\rho d}$.

Although P&D routing involves multiple vehicles, it seems more appropriate to estimate $d(\bar{\rho}_d, C_{\rho d})$ from traveling-salesman (TS) models than vehicle-routing (VR) models. There are two reasons for this: (1) vehicle districts are traversed multiple times during the day, without returning to

the station; and (2) line-haul can occur outside the time window. Each factor reduces the significance of line-haul cost relative to local cost. Unlike VR models, TS models do not include line-haul cost.

Borrowing from TS results, for a homogeneous stop density (i.e., $C_{\rho d} = 0$) and Euclidean metric, Stein (1978) estimated that $d(\bar{\rho}_d, C_{\rho d})$ to asymptotically approaches $.75/\sqrt{\bar{\rho}_d}$. More recently, Johnson (1988) has revised the estimate to $.72/\sqrt{\bar{\rho}_d}$. Both values are somewhat larger than what the VR model predicts: $.57/\sqrt{\bar{\rho}_d}$ (see Daganzo, 1984; Hall, 1990). For the reasons cited above, Johnson's coefficient will be adopted in this paper.

If stop density is randomly distributed across space, with probability density $f(\rho)$, the mean separation for the Euclidean metric would be:

$$d(\bar{\rho}_d, C_{\rho d}) \approx \frac{\int \rho (.72/\sqrt{\rho}) f(\rho) d\rho}{\int \rho f(\rho) d\rho} = \frac{.72 \int \sqrt{\rho} f(\rho) d\rho}{\bar{\rho}_d} . \qquad (2)$$

For example, if stop density is distributed according to a gamma (α,β) distribution (i.e., $\bar{\rho}_d = \alpha/\beta$, $C_{\rho d} = 1/\sqrt{\alpha}$), $d(\bar{\rho}_d, C_{\rho d})$ equals:

$$d(\bar{\rho}_d, C_{\rho d}) = [.72]\frac{\Gamma(\alpha+.5)}{\Gamma(\alpha)\sqrt{\beta}} = [.72\sqrt{\bar{\rho}_d}]\frac{C_{\rho d}\Gamma[1/C^2_{\rho d} + .5]}{\Gamma[1/C^2_{\rho d}]}, \qquad (3)$$

where $\Gamma(\ )$ is the gamma function. Eq. 3 is calculated for several values of α below:

α	1	2	3	4	5	6	∞

$C_{\rho d}$	1	.71	.58	.50	.45	.41	0
$\frac{d(\bar{\rho}_d, C_{\rho d})}{.72\sqrt{\bar{\rho}_d}}$	.89	.94	.96	.97	.98	.98	1.0

As can be predicted from Jensen's inequality, $E(\sqrt{\rho}) \leq \sqrt{E(\bar{\rho})}$, meaning that route length is shorter when stop density is not homogeneous. However, the coefficient of variation must be quite large (.6 or more), before the stochastic solution differs appreciably from the deterministic solution.

As another example, if stop density has a uniform distribution over [a,b] ($\bar{\rho} = (a+b)/2$, $C_{\rho d} = (b-a)/\sqrt{3}(b+a)$), $d(\bar{\rho}, C_{\rho d})$ is defined by:

a/(b-a)	0	.25	.50	.75	1.00	1.50
$C_{\rho d}$	.58	.51	.46	.42	.38	.36
$\frac{d(\bar{\rho}_d, C_{\rho d})}{.72/\sqrt{\bar{\rho}_d}}$	.94	.97	.98	.985	.989	.993

Again, distance is nearly identical to the homogeneous case, except when $C_{\rho d}$ is quite large (.5 or higher).

Taking all factors into account:

$$N_d = (\bar{\rho}_d B)[s_d + (k/\sqrt{\bar{\rho}_d})/v_d] \; / \; W_d \; , \qquad (4)$$

where k is a multiplier specific to the travel metric and probability distribution for ρ. Eq. 4 specifies the minimum number of routes needed to meet mean daily delivery demand. In addition, some safety margin may be

needed to allow for daily variations in workload and stop times. The issue of setting the safety margin will not be addressed in this paper.

PICKUP WORKLOAD

The number of routes required for the afternoon pickups depends on the set of call-in customers remaining to be picked up at the afternoon cutoff time, as well as the set of regular customers. This set does not constitute a uniformly random sample of stops, for it depends on the pickup strategy employed prior to the cutoff. If, for instance, a driver chooses to concentrate on a small section of his pickup region, the spatial distribution of stops remaining at the cutoff will be altered, both in mean and variance. Hence, *the key to minimizing the number of routes is to effectively route vehicles prior to the cutoff*, so that the remaining work is minimized.

The pickup strategy differs from the delivery strategy in that the route must be updated as new calls arrive. In addition, sections of the service region may be covered multiple times to serve newly arriving calls. In these respects, routing pickup vehicles is more complicated than routing delivery vehicles. Therefore, the remainder of the paper is divided into parts, and addresses the following issues in order:

(1) Routing the pickup vehicle between the cutoff time and the end of the pickup time window.
(2) Routing the pickup vehicle prior to the cutoff time, for a homogeneous stop density.
(3) Routing the pickup vehicle prior to the cutoff time, for a stop density that is not homogeneous.

For the sake of simplicity, the paper will assume (realistically) that regular customers can be picked up any time between the cutoff and the end of the workday.

Routing After the Cutoff Time

Once the cutoff is reached, no new calls are allowed and vehicle routing becomes static. Let the following values pertain to the set of stops remaining to be picked up at the cutoff time:

W_p = size of time window available for pickups
N_p = minimum number of routes needed for pickups
$\bar{\rho}_p$ = spatial mean density for pickup stops
s_p = mean time to make a pickup stop
v_p = mean vehicle velocity during pickups.

Then the minimum number of vehicles needed is defined by:

$$N_p = (\bar{\rho}_p A)[s_p + (k/\sqrt{\bar{\rho}_p})/v_p] \; / \; W_p, \qquad (5)$$

where k is a multiplier specific to the distance metric and density distribution.

Pickup Strategy Prior to Cutoff Time: Homogeneous Case

In this section, the stop density of arriving calls is assumed to be homogeneous across space and each stop is assumed to comprise a single

shipment. Under these conditions, there is no reason to favor one stop type over another, or one region over another. Any imbalance in service would cause the mean distance between stops to increase; hence, service rate would decline. Therefore, each vehicle will cover its district in a cyclic path, picking up all new shipments in the path's vicinity as it passes by.

Let:

$\lambda(t)$ = district-wide arrival rate of call-in pickups at time t

$\Lambda(t)$ = cumulative calls to have arrived by time t in a district

$\Omega(t)$ = cumulative stops picked up by time t in a district

A = size of pickup vehicle district (a portion of the entire service region).

At any time t, the number of outstanding calls waiting to be served equals $\Lambda(t)-\Omega(t)$, so the spatial mean stop density is:

$$\bar{\rho} = [\Lambda(t)-\Omega(t)]/A \ . \tag{6}$$

Due to the cyclic nature of the vehicle path, the *outstanding* calls are never uniformly distributed across space. If the system reaches equilibrium, the stop density will vary from 0, in the section of the district just served, up to $2\bar{\rho}$, in the section visited next. Therefore, an approximation for the time to serve a stop is:

t_p = mean time to serve a pickup stop

$$\approx s_p + k/\sqrt{2\bar{\rho}}/v_p = s_p + k/\sqrt{2[\bar{\Lambda}(t)-\bar{\Omega}(t)]}/A/v_p \ , \tag{7}$$

where k ≈ .72 for a Euclidean traveling salesman tour. Dividing k by the factor of $\sqrt{2}$ yields a coefficient of .51. This value is somewhat less than the coefficient of .64 obtained by Bertsimas & Van Ryzin (1990) for a nearest neighbor dynamic traveling salesman tour. These two coefficients can be viewed as lower and upper bounds on the true optimum, the former being based on the idealized assumption that a dynamic route can be served as efficiently as a static route; the latter being based on a non-optimal heuristic.

Eq. 7 translates into a service rate of:

$$\mu(t) \approx \frac{1}{s_p + k/\sqrt{2}[\bar{\lambda}(t)\bar{\Omega}(t)]/\bar{\lambda}/v_p} \tag{8}$$

$$\Omega(t) = \int_0^t \mu(\tau)d\tau \ . \tag{9}$$

In reality, prior to reaching equilibrium the stop density in the vicinity of the vehicle will be less that $2\bar{\rho}$, and can be approximated by:

$$\rho' = \text{stop density in area to be served next}$$
$$\approx (T'/\bar{T})\bar{\rho}, \tag{10}$$

where:

T' = length of time since area to be served next was last served

$\bar{T}$ = spatial average time since area was last served.

Initially, when the vehicle begins its route, T' is zero everywhere (hence T'/T = 1). But as the vehicle proceeds the distribution of T' approaches a uniform [0,2T] distribution, with the maximum in the area to be served next (the minimum in the area served last). Hence, Eq. 8 can initially overestimate $\mu(t)$ by as much as 41%. However, because the initial service rate is very small, this error does not have a significant lasting impact on $\Omega(t)$.

Using Eq. 8, Figure 1 shows the evolution of $\Omega(t)$ for two arrival curves, one stationary, and the other non-stationary. The non-stationary case reflects actual arrival patterns, which sees peaking immediately prior to the cutoff. In both figures, $\mu(t)$ (slope of $\Omega(t)$) begins small, but increases as $\Lambda(t)$-$\Omega(t)$ grows. This is because the driving time per stop declines as the stop density increases. Put another way, drivers are busy throughout the period, even when the arrival rate is low. It is the driver productivity -- not idleness -- that improves as density increases and distance between stops declines.

For the stationary case, equilibrium behavior can be analyzed to measure the impact of system attributes on performance. Equilibrium exists when the service rate equals the arrival rate:

$$\lambda = \mu = \frac{1}{s_p + k/\sqrt{2\tilde{\rho}}/v_p}, \qquad \lambda < 1/s_p , \tag{11}$$

$\tilde{\rho}$ = equilibrium stop density

$$= (1/2)\left[\frac{k/v_p}{1/\lambda - s_p}\right]^2 , \qquad \lambda < 1/s_p . \tag{12}$$

As time progresses, $\bar{\rho}$ approaches $\tilde{\rho}$ from below. The time required to reach equilibrium, τ_0, must be at least as large as the minimum time required to accumulate a queue size of $\tilde{\rho}A$:

$$\tau_0 \geq \left(\frac{\tilde{\rho}A}{\lambda - \mu_{min}}\right) , \tag{13}$$

where,

$$\mu_{min} = \text{the minimum service rate as } \bar{\rho} \text{ approaches zero}$$
$$\approx \frac{1}{s_p + .51\sqrt{A}/v_p} . \tag{14}$$

The approximation for μ_{min} is based on the average distance between two randomly selected points in a square of size A. As Eq. 13 suggests, the time to reach equilibrium is longest when s_p, $\tilde{\rho}$ and A are large, and λ is small. This can be seen by comparing Fig. 2, for which $1/\lambda << s_p$, to Fig. 1, for which $1/\lambda$ is close to s_p.

In addition to the call-in customers, regular customers instantaneously become available for pickup at the cutoff, combining to a mean stop density of $\bar{\rho}_p = \tilde{\rho} + \rho_r$, where ρ_r is the density of regular customer stops. Then the time required to serve all outstanding work at the cutoff is approximately:

$$T_r = A\left[(\tilde{\rho}+\rho_r)s_p + k'\sqrt{(\tilde{\rho}+\rho_r)}\ /v_p\right] . \tag{15}$$

At the cutoff, the probability distribution for ρ varies uniformly between ρ_r and $\rho_r + 2\tilde{\rho}$. If ρ_r=0, with the Euclidean metric, k' is approximately .94×.72,

as shown earlier. For larger values of ρ_r, k' increases toward .72, due to a decline in the coefficient of variation in ρ. For the sake of simplicity, we will assume that k' $\approx$ (.97)(.72) = .70 in our analysis.

T_r cannot exceed the size of the time window, W_p. Substituting Eq. 12 for $\tilde{\rho}$ in Eq. 15, district size is limited by:

$$W_p \geq T_r \tag{16a}$$

$$1 \geq \frac{A'}{\gamma}\left[\frac{A'^2}{2(1-A')^2} + \alpha\gamma + (.97)\sqrt{\frac{A'^2}{2(1-A')^2} + \alpha\gamma'}\right] \tag{16b}$$

where: $A' = \tilde{A\lambda} s_p$

$\alpha = \rho_r/\tilde{\lambda} W_p$

$\gamma = v_p^2 W s_p^2 \tilde{\lambda}/k^2$

$\tilde{\lambda}$ = arrival rate per unit area = λ/A.

A' can be interpreted as the district size, normalized relative to the size of a region that would generate one call in the time s_p. Eq. 16 was solved numerically for A' as a function of α and γ. Results are shown in Figure 3. In no case can A' exceed $\max\{1,1/\alpha\}$, which is the limit of A' as $\gamma \to \infty$. These limits are only attained when driving time is negligible relative to stop time. As γ declines, driving time becomes a more significant factor, which causes district size to decline.

Figure 4 illustrates the relationship between A and the parameters ρ_r and $\tilde{\lambda}$ for the following class of problems:

W_p = 1.5 hours v_p = 20 miles/hour s_p = .05 hours k = .72 .

Level curves are show for values of A ranging from 1.0 square-miles to 2.2 square-miles. The solid lines are the equilibrium results (derived from Figure 3). The dashed lines account for how much time elapsed between the start of the pickup period and the cutoff time, and are based on Eq. 8 (the longer this elapsed time, the closer the result is to equilibrium). As predicted by Eq. 13, the equilibrium result is accurate when the time preceding the cutoff exceeds τ_0. Specifically, Eq. 13 predicts the following:

τ_0 = 2.0 hours when:

A (mi^2)	1.0	1.2	1.4	1.6	1.8	2.0	2.2
$\tilde{\lambda}$ cust/mi^2-hr	16.0	13.2	11.2	9.7	8.6	7.7	6.9

τ_0 = 5.0 hours when:

A (mi^2)	1.0	1.2	1.4	1.6	1.8	2.0	2.2
$\tilde{\lambda}$ cust/mi^2-hr	17.5	14.5	12.3	10.7	9.5	8.5	7.7

Note that these values roughly match the points where the equilibrium curves diverge from the dashed lines.

As a final comparison, Figure 5 demonstrates the relationship between district size and the percentage of customers that are call-in (the remainder are regular, with a combined density of $t_0\lambda + \rho_r$ = 40 customers/$mile^2$). Because call-ins can be served before the cutoff, the allowable district size enlarges as the percentage increases. This is something of a paradox, for it

is inherently more efficient to serve customers when they all arrive at once. In fact, the vehicle miles traveled during pickup are minimized when 100% of the customers are regular. Nevertheless, as the figure shows, spreading the calls over a longer period effectively increases vehicle capacity, even if vehicles have to travel over longer routes.

NON-HOMOGENEOUS STOP DENSITY

With non-homogeneous density, it may be desirable to favor one part of a district over another, either to increase the service rate or to achieve a more favorable stop distribution at the cutoff. In the first two parts of this section, all call-in stops are identical in stop time and each stop generates exactly one shipment, but stop density varies over space. In the third part, the stops are allowed to generate multiple calls in a day.

Identical Stops/Varying Density

As already demonstrated, the time required to serve a set of stops is insensitive to the coefficient of variation of the stop density. Hence, a reasonable heuristic for routing vehicles is to maximize the rate at which stops are served, without regard to the effect on spatial distribution.

To maximize service rate, the vehicle should ideally always serve the location with the largest density of outstanding stops. As a practical matter, separation between regions may prevent the vehicle from fulfilling this goal. However, ignoring the separations allows for an approximate solution that provides some insight into optimal routing. Let:

$\rho(x,y,t)$ = stop density/time of arriving calls, at location (x,y), and time t.

$\lambda(t) = \int_{x,y} \rho(x,y,t)dxdy$.

$q(x,y) = \rho(x,y,t)/\lambda(t)$.

By assumption, q(x,y) will be time invariant.

The density of outstanding stops at any location (x,y) depends on $\rho(x,y,t)$ and the length of time since the location was last served. Therefore, the location with the largest density of outstanding stops is not necessarily the location where q(x,y) is largest. If the vehicle were to spend all of its time in a single region, the density of outstanding stops could fall below that in other regions, with smaller values of q(x,y). Rather, the vehicle should distribute its effort in a manner that equalizes the density of outstanding stops in its vicinity. If the density is not equalized, then the vehicle could spend more time in the area with the highest density of outstanding stops until densities are equalized.

Suppose that the arrival pattern is stationary, and that the system reaches steady-state. Then the cycle time between vehicle visits to any location (x,y) must be proportional to q(x,y). Over this cycle, the stop density must vary between 0, just after the visit, to $2\bar{\rho}$, at the time of the visit, where $\bar{\rho}$ is the spatially invariant mean density of outstanding stops. Therefore:

$$\lambda = \mu = \frac{1}{s_p + k/\sqrt{2\bar{\rho}}/v_p} , \qquad \lambda \leq 1/s_p . \tag{17}$$

Eq. 17 is identical to Eq. 11, the equilibrium condition for homogeneous stop density. The inference here is that variations in spatial stop density should not affect the pickup workload, provided that the driver adopts a location specific cycle time, adjusted according to q(x,y).

A Discretized Case

Suppose now that a vehicle district has just two parts, one with a low stop density and the other with a high stop density. Further, suppose that calls do not arrive continuously over time, but at discrete time points.

Initially, suppose that calls arrive at just two time points, with the second time point interpreted as the cutoff time. After the cutoff, the driver must complete one cycle through the entire region to serve all outstanding calls. The question to answer is how should the driver allocate his effort prior to the cutoff?

Let:

A_i = area of region i (each region is part of a vehicle's district)

ρ_i = stop density in region i, based on arrivals at 1'st time point

λ_j = arrival multiplier, time point j ($\lambda_1 = 1$ by definition)

t_j = time separation between the jth and j+1'st time points

S_j = size of region served between the j'th and j+1'st time points.

First, suppose that the vehicle allocates its entire effort, prior to the cutoff, to a single region, without completing the region. Further suppose,

without loss of generality, that Region 1 is selected. Then the area served would satisfy:

$$S_1\rho_1 = \frac{t_1}{s_p + k/\sqrt{\rho_1}/v_p} \rightarrow S_1 = \frac{t_1/\rho_1}{s_p + k/\sqrt{\rho_1}/v_p} . \quad (18)$$

The time required to serve all remaining work after the cutoff would be:

$$t_2 = S_1(\rho_1\lambda_2)(s_p + k/\sqrt{\rho_1\lambda_2}/v_p) + (A_1 - S_1)[\rho_1(1+\lambda_2)][s_p + k/\sqrt{\rho_1(1+\lambda_2)}/v_p] + A_2\rho_2(1+\lambda_2)[s_p + k/\sqrt{\rho_2(1+\lambda_2)}/v_p] . \quad (19)$$

Substituting Eq. 18 for S_1 in Eq. 19:

$$t_2 = \left[A_1\rho_1(1+\lambda_2)[s_p + k/\sqrt{\rho_1(1+\lambda_2)}/v_p] + A_2\rho_2(1+\lambda_2)[s_p + k/\sqrt{\rho_2(1+\lambda_2)}/v_p]\right] - t_1\left[\frac{s_p + k/\sqrt{\rho_1}/v_p[\sqrt{1+\lambda_2} - \sqrt{\lambda_2}]}{s_p + k/\sqrt{\rho_1}/v_p}\right] . \quad (20)$$

The first component of Eq. 20 equals the total time to process all work, if all routing were performed after the cutoff. The second term is the work reduction due to routing prior to the cutoff. Only when $\lambda_2 = 0$, or when $\rho_1 \rightarrow \infty$, can the work reduction equal t_1, the actual time available. Otherwise, the work reduction is less than t_1, the difference arising from the lower productivity when serving only a portion of the total calls that eventually arrive in S_1.

The work reduction in Eq. 20 is largest when ρ_1 is maximized, meaning

that the vehicle should serve the region with the largest stop density. Doing this will minimize t_2 or, if t_2 is a fixed time window, allow the total district size (A_1+A_2) to be maximized. Maximizing district size fulfills the objective of minimizing the number of routes. (If it were possible to complete an entire region prior to the cutoff, then that region should also clearly be the one with the larger density.)

To carry the discrete case further, a vehicle can be routed within its district by traveling from region to region, with the objective of minimizing outstanding work at the cutoff time.

$$W(t_J) = \text{quantity of outstanding work, at cutoff time } t_J$$
$$= \sum_{i=1}^{I} A_i \hat{\rho}_i [s_p + k/\sqrt{\rho}/v_p] \ , \qquad (21)$$

where, $\hat{\rho}_i$ = outstanding stop density in region i

I = number of regions.

The outstanding stop density is defined by the routing sequence prior to the cutoff time, which can be viewed as a sequence of time periods, each of which represents a single cycle through a region. Each cycle has duration:

$$t_j = A_i \bar{\rho}_i [s_p + k/\sqrt{\bar{\rho}_i}/v_p] + T\{i(j-1),i\} \ , \qquad (22)$$

where, $T\{i(j-1),i\}$ = time to travel from region visited in previous period, $i(j-1)$, to region i.

The optimization problem could be formulated and solved as a dynamic program, though it seems that a reasonable heuristic would be to "select the region with the largest outstanding stop density" next (perhaps modified to reflect distances between regions).

Non-Identical Stops

Stops can differ according to the number of pieces shipped and according to the likelihood that a stop would generate more than one call during a day. The stop time, s_p, is composed of an access/egress time, a time greeting the customer, and a time processing the pieces.

$$s_p = \nu_p + \kappa \cdot n \ , \tag{23}$$

where:

ν_p = access/egress time + greeting time

κ = time to process a piece

n = number of pieces shipped.

As indicated in the prior sections, the goal prior to the cutoff time should be to maximize driver productivity. Eq. 20 provides a framework for measuring productivity. That is, the productivity is the proportionate reduction in work after the cutoff, due to work completed prior to the cutoff. Productivity losses stem from two sources: (1) added distance between stops, due to a lower stop density, and (2) repetition of stops due to multiple

calls. Put another way, prior to the cutoff, the driver is only fully productive during: (A) access/egress/greeting at stops that only generate a single call per day, and (B) processing individual pieces.

Taking these factors into consideration, hybrid routing strategies might include:

(1) Favoring calls that contain many pieces, to increase the proportion of time spent processing pieces.
(2) Skipping stops that are likely to generate multiple calls, to prevent the access/egress and driving time from being wasted.

Because the piece processing time tends to be a small percentage of the total, the first strategy is really not very practical. Further, it conflicts with the second objective, because stops that generate many pieces are also likely to generate many calls. Therefore, the first possibility will not be considered.

The second strategy might be implemented by transferring some customers from the call-in category to the regular category; that is, by waiting until the cutoff time before visiting them. In terms of the models presented, the net effect will be to reduce $\tilde{\lambda}$ and increase ρ_r. The question, then, is whether such a change increases or decreases the quantity of work remaining to be completed at the cutoff.

For any pickup system, this question can be addressed by constructing a trade-off curve, as in Figure 4. Let:

t_0 = length of pickup period preceding cutoff

$\tilde{\lambda}_0$ = rate at which selected class of customers generates calls

ρ_0 = stop density for selected customer class.

If the selected class is moved from the call-in category to the regular category, $\tilde{\lambda}$ will decrease by $\tilde{\lambda}_0$ and ρ_r will increase by ρ_0. However, for the change to be contemplated, $\rho_0/\tilde{\lambda}_0 < t_0$; otherwise, the customer class would not be generating multiple calls, on average.

To determine whether a transfer reduces the workload, the ratio $\rho_0/\tilde{\lambda}$ can be compared against the marginal rate of substitution between ρ_r and $\tilde{\lambda}$, as defined by the level curves for district size. Specifically, a transfer will result in marginal work reduction when:

$$\rho_0/\tilde{\lambda}_0 < -(\partial\rho/\partial\lambda) \qquad (\rho_0/\tilde{\lambda}_0 < t_0) \ . \tag{24}$$

For example, the following data is taken from Figure 4 with $\rho_r = 10$:

t_0 = 2.0 hrs.			t_0 = 5.0 hrs.		
$\tilde{\lambda}$	A	$-\partial\rho/\partial\tilde{\lambda}$	$\tilde{\lambda}$	A	$-\partial\rho/\partial\tilde{\lambda}$
19	1.2	1.8	16	1.2	4.8
15	1.4	1.6	13	1.4	3.9
12	1.6	1.5	11	1.6	3.2
10	1.8	1.2	9	1.8	2.6
8	2.0	.6	7.5	2.0	1.5
6.5	2.2	.4	6.0	2.2	1.0

To interpret the data, transferring stops from the call-in category to the regular category is most attractive when $\tilde{\lambda}$ is large, in which case the vehicle has a heavy workload prior to the cutoff. On the other hand, if $\tilde{\lambda}$ is small, a transfer may be unattractive -- even if it means saving repeated stops. For instance, if $\tilde{\lambda}$ is 6.5, $\rho_0/\tilde{\lambda}_0$ can be as small as 1/5 t_0 (i.e., each stop generates 5 calls prior to the cutoff) and it would still be preferable to serve the stops as call-ins.

As a caveat, the preceding analysis assumes that each call generates exactly one vehicle visit. In fact, if the cycle time between pickups is sufficiently long, multiple calls might be made between visits. Hence, the marginal rate of substitution will overestimate the benefit of shifting customers from the call-in to the regular category in extreme cases.

CONCLUSIONS

This paper has demonstrated how the constraints of overnight delivery affect the design of pickup and delivery systems. For such systems, cost depends on the number of vehicle routes needed to pickup and deliver the shipments. This number of routes depends on the critical parts of the driver's day: the morning delivery period, up to the delivery deadline, and the afternoon pickup period, after the cutoff time. Of the two, pickup routing is the more complicated, due to the dynamic nature of customer calls. In pickup routing, the goal is to minimize the workload of outstanding calls at the cutoff time. Meeting this goal depends on how regions are sequenced for pickup prior to the cutoff. A reasonable heuristic seems to be the

"highest stop density next" rule.

Although the paper did not yield a routing algorithm, it did provide approximate models that can be used to estimate the effect of changes in customer base on cost. In particular, the pickup model can be compared to the delivery model to identify the dominant period of the day. Even though pickup routing is less efficient (due to dynamic calls), and even though it takes longer to process a pickup than a delivery ($s_p > s_d$), it is not necessarily dominant. Mainly, this is because pickup stops tend to generate more pieces than delivery stops. Put another way, delivery routes tend to contain more stops than pickup routes.

One application of the models may be to adjust delivery and cutoff times to attain a balanced workload. Another application would be to measure the change in cost due to changes in service standards. Unfortunately, these analyses are not straight-forward because any change in service standards will surely affect the demand pattern. Sensitivity analysis, combined with some form of demand modeling, is needed to answer these strategic questions.

Finally, the models presented do no account for all of the phenomena presented at the end of the "system description" section. Excluded factors include (1) generation of calls at different locations within a building and (2) design of special routes dedicated to serving large customers. These topics could be the basis for future research.

REFERENCES

Bertsimas, D.J. and G. van Ryzin (1990). "A stochastic and dynamic vehicle routing problem in the euclidean plane," MIT Operations Research Center, working paper OR 210-90.

Chan, Y. and R.J. Ponder (1979). "The small package air freight industry in the United States: a review of the Federal Express experience," Transportation Research, 13A, 221-229.

Chestler, L. (1985). "Overnight air express: spatial pattern, competition and the future of small package delivery services," Transportation Quarterly, 39, 59-71.

Daganzo, C.F. (1978). "An approximate model of many-to-many demand responsive transportation systems." Transportation Research, 12, 325-333.

Daganzo, C.F. (1984). "The distance traveled to visit N points with a maximum of C stops per vehicle: an analytical model and an application." Transportation Science, 18, 331-350.

Daganzo, C.F. (1987). "Modeling distribution problems with time windows: parts I and II." Transportation Science, 21, 171-187.

Emery, J.C., A.J. Jones, J.W. Alden, J.M. Dauernheim (1986). "The revolution in the air courier/express industry." Proceedings of the 13th International Forum on Air Cargo, Basel, Switzerland, 91-107.

Finnegan, W.F. and Andrade, J.M. (1984) "The impact of changes in the environment on an airline network." AGIFORS 24th International Symposium, Proceedings, Strasbourg, France.

Hall, R.W. (1987) "Comparison of strategies for routing shipments through transportation terminals." Transportation Research, 21A, 421-429.

Hall, R.W. (1989) "Configuration of an overnight package air network," Transportation Research, 23A, 139-149.

Hall, R.W. (1991) "Design of Local Area Freight Networks," submitted to Transportation Research.

Hall, R.W., Du, Y. and Lin, J. (1990). "Integration of discrete and continuous models for routing vehicles," submitted to Transportation Science.

Jaillet, P. (1988) "A priori solution of a traveling salesman problem in which a random subset of the customers are visited." Operations Research, 36, 929-936.

Johnson, D. (1988), talk presented at the Mathematical Programming Symposium, Tokyo.

Kanafani, A. and Ghobrial, A.A. (1985) "Airline hubbing-some implications for airport economics," Transportation Research, 19A, 15-27.

O'Kelly, M.E. (1986) "The location of interacting hub facilities," Transportation Science, 20, 92-106.

Psaraftis, H. (1988) "Dynamic vehicle routing problems," in Vehicle Routing: Methods and Studies (B. Golden and A. Assad, eds.), New York: North Holland.

Stein, D. (1978) "An asymptotic probabilistic analysis of a routing problem." Mathematics of Operations Research, 3, 89-101.

Stein, D. (1978) "Scheduling dial-a-ride transportation systems." Transportation Science, 12, 232-249.

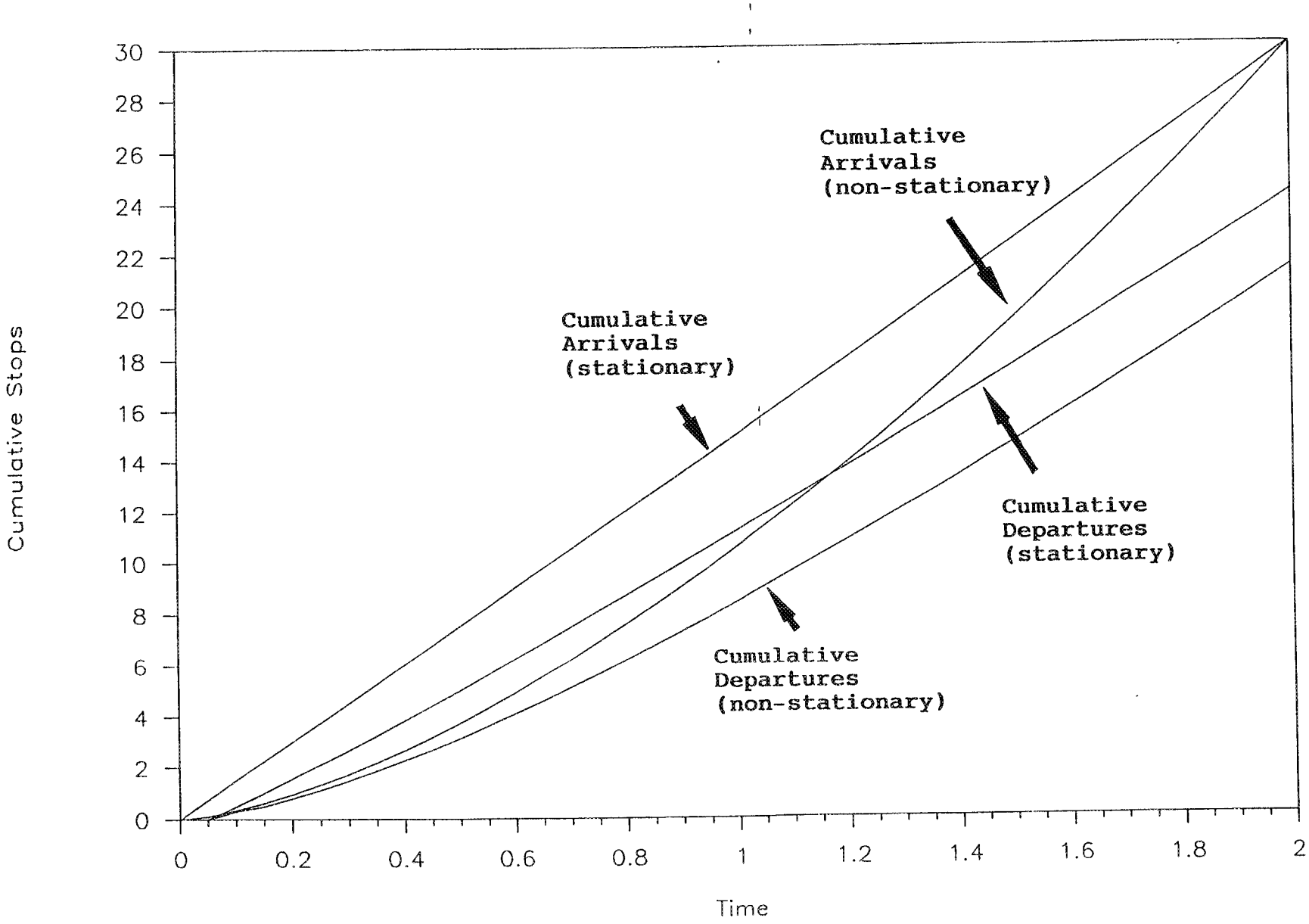

Figure 1. Evolution of cumulative departure curves for stationary and non-stationary arrival patterns.

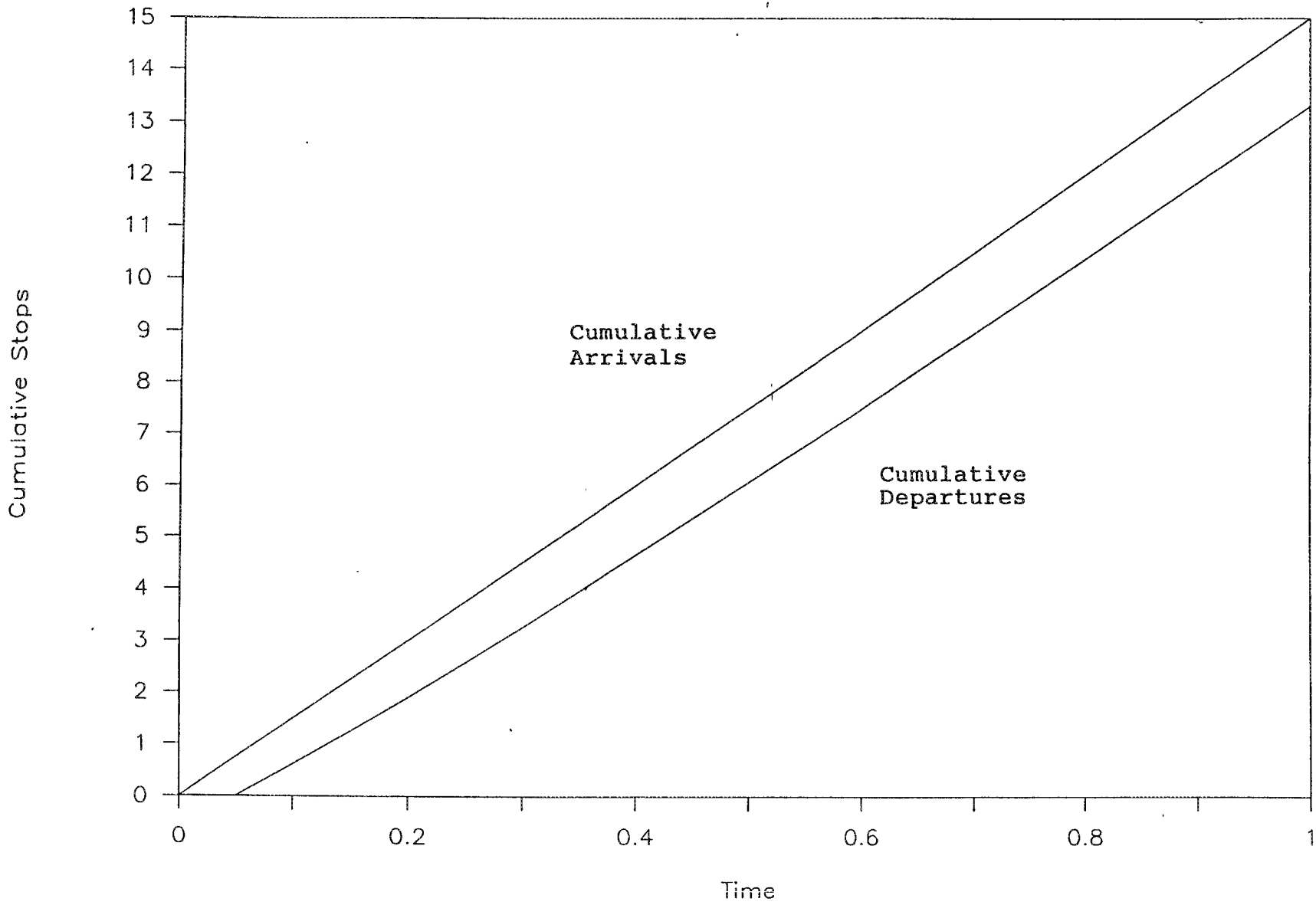

Figure 2. Evolution of cumulative departure curve for system that quickly approaches equilibrium

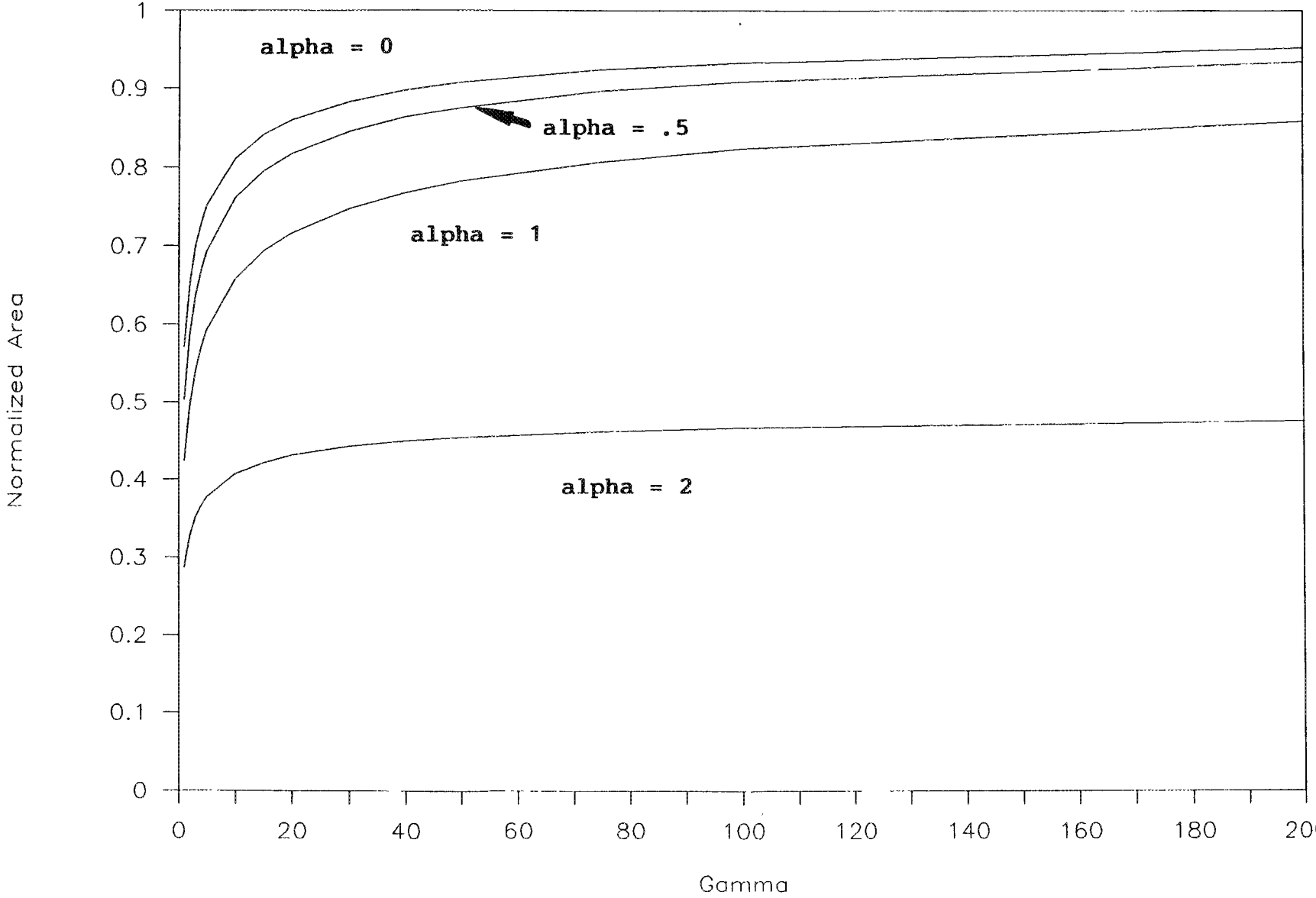

Figure 3. Normalized district size increases as gamma increases and alpha decreases

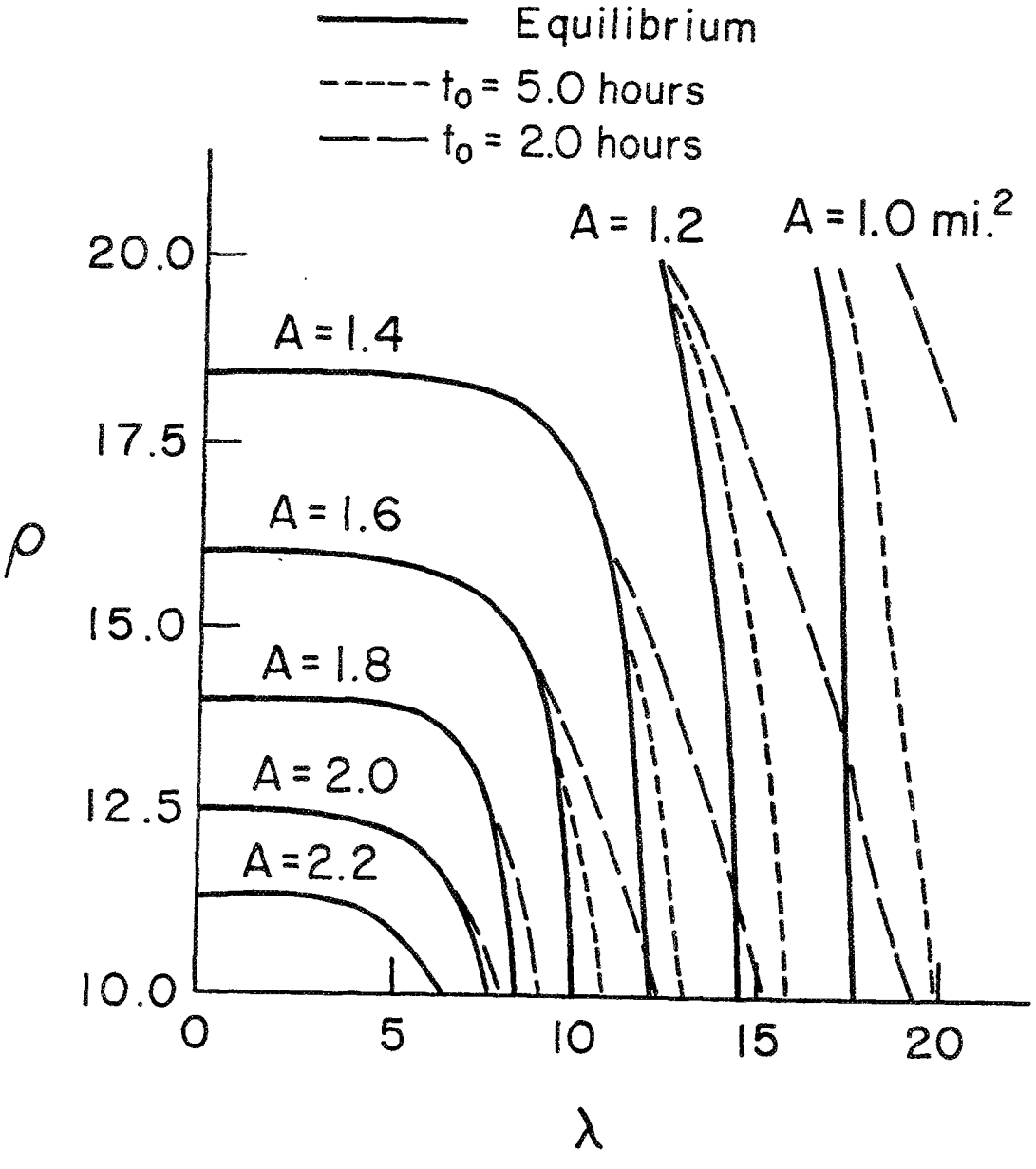

Figure 4. Level curves for district size show influence of number of regular customers/area (rho) and arrival rate of call-in customers (lambda). Equilibrium result is most accurate for small t_0 and underloaded system. (W_p = 1.5 hrs., v_p = 20 miles/hour, s_p = .05 hrs., k = .72)

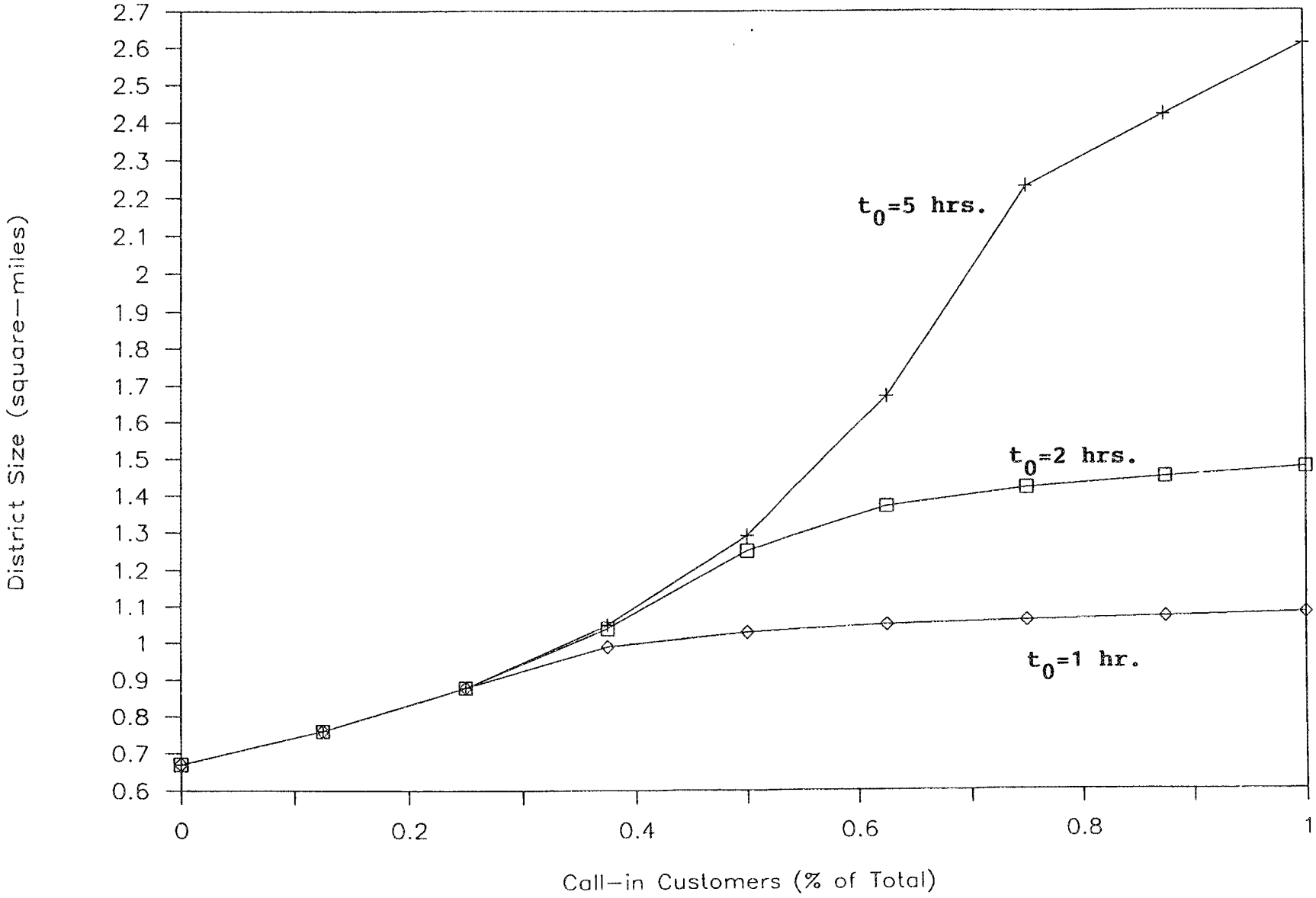

Figure 5. Vehicles can serve larger districts when a large proportion of customers are call-in (W_p=1.5 hrs., v_p=20 mph, s_p=.05 hrs.,k=.72)