Cost Modeling and Design Techniques for Integrated Package Distribution Systems

Karen R. Smilowitz* and Carlos F. Daganzo†

October 15, 2003

Abstract

Complex package distribution systems are designed using idealizations of network geometries, operating costs, demand and customer distributions, and routing patterns. The goal is to find simple, yet realistic, guidelines to design and operate a network integrated both by transportation mode and service level; i.e., overnight (express) and longer (deferred) deadlines. The decision variables and parameters that define the problem are presented along with the models to approximate total operating cost. The design problem is then reduced to a series of optimization subproblems that can be solved easily. The proposed approach provides valuable insight for the design and operation of integrated package distribution systems. Qualitative conclusions suggest that benefits of integration are greater when deferred demand exceeds express demand. This insight helps to explain the different business strategies of package delivery firms today.

This paper introduces design strategies and cost modeling techniques for multiple mode, multiple service level package delivery networks where service levels are defined by guaranteed delivery times (i.e., overnight, two-day delivery). Such research is critical at a time when new technology

*Department of Industrial Engineering and Management Sciences, Northwestern University
†Department of Civil and Environmental Engineering, University of California, Berkeley
and the global economy are revolutionizing freight transportation. It is important to understand how companies adapt to these changes. For example, transportation providers now offer a wider range of service levels to increase market share and utilize resources more efficiently. New network configurations and routing strategies are possible when one considers integration of service levels and transportation modes.

The design and operation of large-scale transportation networks are difficult due to the number of decision variables and constraints, and their intricate interdependencies. This is particularly true for the complex hierarchical networks adopted for package delivery. Unlike passengers in air networks, shipments in freight networks can be routed in more circuitous ways to achieve economies of scale and density, provided time constraints are not violated. Conventional network design and routing models cannot sufficiently capture the complexity of multimode, multiservice networks. This paper examines various degrees of mode and service level integration for package delivery and presents a complete modeling framework for strategic design problems for large-scale integrated distribution networks. While the network design problem is quite complex, we demonstrate the ability to estimate costs and obtain designs for such systems using continuum approximations. These estimates are used to evaluate potential mergers of package distribution systems.

Section 1 discusses network configurations and routing principles for package delivery. Section 2 investigates both discrete and continuous formulations of the network design problem. Section 3 presents approximation methods for the network design problem and Section 4 the solution method. Section 5 presents a case study. Finally, Section 6 summarizes the research.

1 Integrated Package Distribution Networks

Package delivery firms operate very complex networks. A typical Federal Express or UPS package passes through a hierarchy of terminals en route from origin to destination, transported by several
modes. Here we present a stylized version of package delivery networks. Two service levels are assumed: express and deferred demand. Express items are highly time sensitive; deferred items are not. Two transportation modes are assumed: air and ground. Local and access (regional) transportation is conducted by ground vehicles (delivery vans, trucks, etc.), but long haul transportation can be performed by ground (tractor-trailers) and air. In integrated delivery networks, express items are transported by air for long haul trips due to tight time constraints.\footnote{Express items with nearby destinations may not travel by air. Such items are ignored in this study.} Deferred items may be sent by ground or air.

![Integrated package distribution network](image)

**Figure 1: Integrated package distribution network**

An integrated distribution network, shown in Figure 1, operates as follows. Items travel via local pick-up tours to the nearest regional consolidation terminal where items within the region are consolidated for efficient long haul transportation. Items are then sent along access routes from consolidation terminals to either breakbulk terminals or airports, depending on the long haul mode.

Items traveling by air are delivered to the nearest airport for an evening flight to the main hub. Aircraft may stop at a second airport to/from the hub to increase aircraft loads and maintain
daily frequencies. Items typically arrive at the main hub between 10 pm and 2 am, where they are sorted by destination airport and loaded onto aircraft for morning departures. After arriving at the destination airport, the ground process is reversed: items travel to a consolidation terminal and then to their final destination.

The long haul ground system includes several breakbulk terminals, which, like airports, act as gateways to the long haul network. However, unlike the air network, there is no single main hub. All breakbulk terminals serve as hubs, albeit for smaller percentages of the total network volume. Items are routed between consolidation terminals through two breakbulk terminals.

We consider three distribution strategies. In all cases, the network design problem determines ground and air network configurations (number and location of terminals) and routing guidelines for both items and vehicles.

**Fully integrated networks: integrated facilities and routing**

Express items are sent by air; deferred items travel either way. Consolidation terminals serve both service types jointly. Savings in local transportation are possible through economies of density. Flexible routing allows excess aircraft capacity to be filled with deferred items, reducing ground transportation needs.

**Semi-integrated networks: integrated facilities, segregated routing**

Routing is performed separately for the two service types at all levels, but consolidation terminals are shared by all service types. Here fewer consolidation terminals can provide the same coverage.

**Non-integrated networks: segregated facilities, segregated routing**

Non-integrated networks are simply the superposition of two separate networks offering express and deferred service independently. Consolidation terminals are mode-specific.
1.1 Related Literature

Two principal approaches have been employed in the literature to address components of this problem: mixed-integer programming with detailed discrete data and continuum approximations. While the former provide a higher level of detail, the latter are more revealing of “the big picture”.

Numerical optimization approaches to network modeling have been studied extensively; see Magnanti and Wong (1984); Ahuja et al. (1993); Ball et al. (1995). As discussed in these and other more general references; e.g., Nemhauser and Wolsey (1999), optimal solutions can be found numerically for small network problems. In some special cases, it is possible to solve large problems. In general, however, as the network size increases, problems become more difficult to solve, and heuristic approaches are often necessary. Numerical optimization models have been successful in solving tactical and operational problems for transportation networks, offering detailed, cost-minimizing operating plans; see review in Crainic (2000), as well as Powell and Sheffi (1983); Barnhart and Schneur (1996); Armacost (2000). However, collecting demand and cost data for these models can be time-consuming and, at times, impossible. These difficulties are compounded if demand is uncertain.

On the other hand, continuum approximation models use smooth functions to describe the data, such as a demand density function that varies with location; see for example Daganzo and Newell (1986). Smooth functions are also used to describe decisions (in place of decision variables), e.g. as in the case of spatially varying terminal densities. Knowledge of these decision functions gives enough information to develop a network configuration and an operating plan with a predictable cost even with uncertain demand; see Daganzo (1999). Early work on approximation methods (Eilon et al. (1971); Newell (1973); Geoffrion (1976)) recognized that approximations provide near optimal solutions and offer valuable insight into operating strategies and network design. Whereas numerical optimization models perform better on smaller problems, the opposite is true with continuum
approximations. The larger the problem, the more accurate the approximations become; see Daskin (1985); Campbell (1993); Daganzo (1999). Continuum formulations sometimes can be decomposed into smaller components that allow the complete solution space to be explored systematically.

Although continuum approximation methods are well suited for the design of large-scale transportation networks, the topic has not been explored thoroughly. This paper fills the following methodological gaps identified in a review paper by Langevin et al. (1996): (i) current multiple origin/multiple destination distribution models do not adequately incorporate multiple transshipments and multistop (peddling) tours; (ii) current models ignore additional operating costs beyond transportation and inventory; (iii) current models have not considered the cost of repositioning empty vehicles, except perhaps Jordan and Burns (1984); Hall (1991); and (iv) current models do not consider multiple service levels although several studies have considered distribution of time sensitive items, see Han (1984); Daganzo (1987a, b); Kiesling (1995). Multiple transportation modes have been included in a limited number of models, see Hall (1989). This paper demonstrates, as is stressed in the continuum literature, that numerical optimization and continuum methods can and should be used together. Continuum approximations are ideally suited for planning purposes, when demand forecasts are uncertain and aggregate. They can suggest system configurations, even before precise data are available. Once detailed data become available, operational details can be further researched with discrete optimization.

2 Network Design Problem

The network design problem minimizes expected transportation costs (fixed vehicle costs and variable operating costs) and facility costs (fixed terminal charges, handling costs, and storage expenses) over a planning horizon while meeting service level constraints. This problem is part of a two-phase approach in which the network is designed first and then operating plans are developed for the fixed
network. As discussed in Daganzo and Erera (1999), the performance of the overall distribution system can be improved if the network is designed with the subsequent operating plans in mind. To clarify ideas and illustrate the complexity of the problem, Section 2.1 presents a discrete formulation using the simplest of the three distribution strategies described in Section 1 (non-integrated networks) and focusing on the ground network for deferred items only.\footnote{A single mode air network can be modeled in a similar manner.} Even without integration the problem is very complicated. Section 2.2 introduces continuous functions to transform the formulation.

### 2.1 Discrete Formulation

A typical path of an item from origin to destination is shown in Figure 2(a). As explained in Section 1, items travel from an origin, to a consolidation terminal, to the long haul network via an airport or breakbulk terminals, and then the process is reversed on the way to the final destination. Since no step is skipped in this hierarchical scheme, operating costs can be separated by distribution level and terminals visited. The set of distribution levels is: $\mathcal{L} = \{0, 1, 2\}$: local (0), access (1) and long haul (2). Level 0 facilities are origins and destinations; level 1 facilities are consolidation terminals; and level 2 facilities are breakbulk terminals in ground networks, and airports and the main air hub in air networks. Let $\mathcal{T} = \{C, B, P, H\}$ denote the set of terminals, consisting of consolidation terminals (C), breakbulk terminals (B), airports (P), and main air hub (H). It is assumed that all vehicles serving a distribution level $l$ have a capacity $V_l$; see Figure 2(b).

Assume $I_l$ possible locations for a terminal of level $l = 1, 2$. Further, there are $I^0$ discrete locations for customers (origins and destinations) in the service area over a planning horizon of $T$ days. An origin-destination pair is defined by origin location and time $(i, r)$ and destination location and due date $(j, t)$ for nodes $i, j \in 1..I^0$ and discrete time instances $r, t \in 1..T$. The demands across
days and origin-destination pairs are assumed to be random and mutually independent. The mean and variance for an origin-destination pair are $\Lambda_{i,r,j,t}$ and $\Omega_{i,r,j,t}$, respectively.

The decision variables in the network design problem are:

$$X_i^l = \begin{cases} 
1, & \text{if a level } l \text{ terminal is opened at location } i = 1..I^l \text{ for } l = 1, 2 \\
0, & \text{otherwise}
\end{cases}$$

$$Z_{i,j}^l = \begin{cases} 
1, & \text{if a level } l \text{ facility at } i = 1..I^l \text{ is served by a level } l+1 \text{ terminal at } j = 1..I^{l+1} \text{ for } l = 0, 1 \\
0, & \text{otherwise}
\end{cases}$$

$$X_i^l = \text{number of vehicles of capacity } V^l \text{ assigned to a terminal at } i = 1..I^1 \text{ for } l = 0; i = 1..I^2 \text{ for } l = 1, 2.$$
can be decomposed by distribution level. Due to day-to-day variations in demand, vehicle routings may change over the planning horizon, but the assignment of vehicles to terminals is assumed to be constant. Therefore, the objective function minimizes the following:

$$\min \sum_{l \in L} \Psi_l (Y, Z) + \sum_{l \in L} \Phi_l (X, Y, Z)$$

(1a)

The first set of constraints (1b)-(1d) defines the network hierarchy, including terminal locations, linkages and customer allocations. Constraints (1b) ensure that every origin and destination is allocated to exactly one consolidation terminal. Constraints (1c) ensure that a consolidation terminal opened at location $i$ is allocated to exactly one breakbulk terminal. Constraints (1d) ensure that only open terminals can serve a lower distribution level.

$$\sum_{j=1..I^1} Z_{i,j}^0 = 1 \quad \forall i = 1..I^0$$

(1b)

$$\sum_{j=1..I^2} Z_{i,j}^1 = Y_i^1 \quad \forall i = 1..I^1$$

(1c)

$$Z_{i,j}^{l-1} \leq Y_j^l \quad \forall i = 1..I^{l-1}, j = 1..I^l, l = 1,2$$

(1d)

Constraints (1e)-(1h) establish the minimum number of vehicles needed to serve local and access tours from each terminal. Let auxiliary variables $W_p^l$ define a daily demand threshold unlikely to be exceeded during the planning horizon from a terminal at $p = 1..I^{l+1}$ for levels $l = 0, 1$. Constraints (1e) and (1f) define $W_p^0$ as the maximum of inbound and outbound daily demand at a consolidation terminal across all days. The last term is a buffer factor that accounts for the standard deviation of the combined demand over origins/destinations served by that terminal on a particular day. The parameter $\alpha$, dependent of the service level of the network, specifies the necessary buffer. It is assumed that demand is transported from consolidation terminals to customers just in time for

---

4Symmetry in distribution is assumed: if a consolidation terminal serves a customer for pick-up, it also serves that customer for delivery.
Constraints (1g) define the threshold for each breakbulk terminal as the aggregation of the thresholds of the consolidation terminals assigned to it. Constraints (1h) ensure that sufficient vehicles are provided to cover the thresholds, assuming items of equal size.

\[
W_{p}^{0} \geq \sum_{i=1..I^0} \sum_{r>t} \sum_{j=1..I^0} Z_{j,p}^{0} \Lambda_{j,t,i,r} + \alpha \sqrt{\sum_{i=1..I^0} \sum_{r>t} \sum_{j=1..I^0} Z_{j,p}^{0} \Omega_{j,t,i,r}} \quad \forall p = 1..I^1, t = 1..T \quad (1e)
\]

\[
W_{p}^{0} \geq \sum_{i=1..I^0} \sum_{r<t} \sum_{j=1..I^0} Z_{j,p}^{0} \Lambda_{i,r,j,t} + \alpha \sqrt{\sum_{i=1..I^0} \sum_{r<t} \sum_{j=1..I^0} Z_{j,p}^{0} \Omega_{i,r,j,t}} \quad \forall p = 1..I^1, t = 1..T \quad (1f)
\]

\[
W_{p}^{1} \geq \sum_{i \in I^1} W_{i}^{0} Z_{i,p}^{1} \quad \forall p = 1..I^2 \quad (1g)
\]

\[
X_{i}^{l} \geq W_{i}^{l} / V^{l} \quad \forall i = 1..I^{l+1}, l = 0, 1 \quad (1h)
\]

Constraints (1i) bound from below the number of long haul vehicles. This is more difficult to write in a simple form since vehicles are not required to return to their bases at the conclusion of delivery. The lower bounds, however, should depend on the breakbulk terminal locations, demand allocations and vehicle capacity; i.e., the constraints should be of the form:

\[
X_{i}^{2} \geq \Gamma(Y^{2}, Z^{1}, V^{2}) \quad \forall i = 1..I^{2} \quad (1i)
\]

Of course, vehicle variables must satisfy integrality constraints (1j), facility location and allocation variables must be binary (1k) and (1l), and \( W_{i}^{l} \) must be non-negative (1m).

\[
X_{i}^{l} \geq 0, \text{ integer} \quad \forall i = 1..I^{l} \text{ for } l = 0; \ i = 1..I^{2} \text{ for } l = 1, 2 \quad (1j)
\]

\[
Y_{i}^{l} \in \{0, 1\} \quad \forall i = 1..I^{l}, l = 1, 2 \quad (1k)
\]

\[
Z_{i,j}^{l} \in \{0, 1\} \quad \forall i = 1..I^{l}, j = 1..I^{l+1}, l = 0, 1 \quad (1l)
\]

\[
W_{i}^{l} \geq 0 \quad \forall i = 1..I^{l+1}, l = 0, 1. \quad (1m)
\]

\(^5\)Although it is possible for items to arrive before their due date, doing so limits the ability of carriers to differentiate service types for marketing purposes.
There are several difficulties with formulation (1). While it is possible to quantify facility costs (Ψ), it is considerably more difficult to quantify transportation costs (Φ) and operating constraints (Γ). The expressions should be sensitive to terminal locations and demand allocations, recognizing the different ways in which vehicles may be routed for each network configuration. Moreover, even if one is successful in this endeavor, heuristics are needed to solve realistic problem instances with hundreds or thousands of possible physical locations. The need to model various levels of network integration complicates the problem even further. Fortunately many of these difficulties can be overcome by a continuum approximation, since, as we shall see, continuity allows one to decompose the problem geographically.

2.2 Continuum approximation: notation and assumptions

Figure 3 shows the mapping of the discrete formulation to a continuum formulation and the reverse mapping to obtain discrete solutions. This section describes the top part of the figure, yet the forecasts of a planning process often produce aggregate demand rates (and other parameters) in continuum form. Then one only needs to express the decision variables continuously without doing anything to the input data. Sections 3 - 5 describe the right box and Section 6 describes the bottom.

The demand, service level, and cost data over a service area $\mathcal{A}$ are defined as functions, which we call parameters, that vary with the coordinates, $x$, of points on the plane. The solution is described in terms of decision functions of location, which we call variables. We now return to multiple modes and multiple service levels. Let $\mathcal{S} = \{E, D\}$ denote the set of service levels for express and deferred demand, and the labels $A$ and $G$ (air and ground) network type. The air network is the set of links and terminals used by items traveling by air, including the ground portion of their travel. The ground network is the set of links and terminals used by items not traveling by air.

---

$^6$Because much notation is needed, a summary is provided in the appendix.
Customer data

We consider stationary, deterministic demand data. Extensions to stochastic demand are simple and shown in Section 3.7; time-dependent demand is discussed in Smilowitz (2001). Exact customer location data for the $I^0$ discrete customers are replaced with spatial customer densities for service level, $\delta^s(x)$ (customers/unit area). The number of points in a subregion $A'$ of $A$ is

$$N_{A'} = \int_{x \in A'} \delta^s(x) dx$$

If $A'$ is chosen such that $\delta^s(x)$ is nearly constant over the subregion, then the number of points is:

$$N_{A'} \approx \delta^s(x)|A'|$$

Origin-destination demands $\Lambda_{i,r,j,t}$ are replaced with temporal demand rates $\lambda^s(x^o, x^i)$ from a region of unit area about $x^o$ to a region of unit area about $x^i$ for service level $s \in S$ (items/area^2*time). Density and demand estimates are natural outputs of a demand forecasting process and can often be obtained from aggregated data; e.g., as shown in Clarens and Hurdle (1975).
Decision variables

Discrete location variables \((Y)\) are replaced with continuous variables that specify the density of terminals of type \(y \in T\) within a region, \(\Delta_y(x)\) \((\text{terminals/unit area})\). Additional decision variables describe item and vehicle routings, in place of \((X, Z)\). Indices \(B = \{i, o\}\) define trips inbound to and outbound from a terminal, and \(V = \{a, t\}\) define vehicle type (air and truck).\(^7\) A set of variables \(\{h_{l}^{m,b}(x)\}\) defines the route headways for distribution level \(l \in L\) for network \(m = A, G\) in direction \(b \in B\). Headways are time intervals between consecutive dispatches; as such they indicate how often a route is run. A set \(\{n_{l}^{m,b}(x)\}\) defines the number of stops on a route; \(\{v_{l}^{m,b}(x)\}\) defines the shipment size for each stop; and \(\{r_{l}^{m}(x)\}\) defines the average linehaul distance for the route.

Service level constraints

A series of service level parameters enforces express and deferred deadlines. The maximum headway length for a route of type \(l \in L\) is \(H_{l}^{m}\) for \(m = A, G\). Tight restrictions on express item delivery are enforced by assuming that \(H_{l}^{A} \leq 1\) day. We limit the number of stops on a route of type \(l \in L\) by introducing an upper bound \(N_{l}^{m}\). A maximum airport service radius, \(\rho\), is introduced to meet air time restrictions.

Additional assumptions

Sorting costs at facilities are approximated as constants independent of all decision variables, as shown in Smilowitz (2001). In-vehicle inventory costs are not considered; therefore, ground vehicles make as many stops as allowed by \(N_{l}^{m}\) and vehicle capacity \(V_{l}^{m}\); see the full vehicle theorem in Daganzo (1999). Therefore ground vehicles will reach either their physical capacity or their maximum number of stops. In all scenarios, it is assumed that the long haul air network is optimally

---

\(^7\)In the discussion here, one truck size is assumed for simplicity. Numerical results in Section 5 consider three truck sizes: the largest trucks for long haul routes, smaller trucks for access routes, and delivery vans for local routes.
configured for express items only (i.e., airport locations are determined by express demand only which, in turn, specifies available excess capacity). It is further assumed that in fully integrated networks as many deferred items as possible are shifted to air. This is reasonable for large ground networks, as confirmed in Smilowitz (2001). The fractional shift in direction \( b \in B \) is \( \omega_b(x) \). It is assumed that each terminal is centrally located within an approximate circular service region; therefore, \( r_l^m(x) \) is approximated as \( 2/3 \) of the radius of the circular region.

\[
r_l^m(x) \approx \frac{2}{3} \left( \frac{\pi}{\Delta_l(x)} \right)^{-\frac{1}{2}}
\]  

(2)

This estimate is on the low side, but quite accurate for desirable terminal arrangements.

3 Continuum approximation: logistic cost functions

Approximations of facility and transportation costs (\( \Psi \) and \( \Phi \)) are decomposed by distribution level as in (1). For a specific pair of origin-destination regions, the average cost per item per unit time, \( z \), is comprised of the following components:

\[
z = z_{\text{local}} + z_{\text{access}} + z_{\text{longhaul}} + z_{\text{reposition}} + z_{\text{CT}} + z_{\text{airport}} + z_{\text{BBT}} + z_{\text{hub}}
\]  

(3)

Equation (3) contains transportation costs for each distribution level: \( z_{\text{local}} \), \( z_{\text{access}} \), and \( z_{\text{longhaul}} \); terminal costs: \( z_{\text{CT}} \), \( z_{\text{airport}} \), \( z_{\text{BBT}} \), and \( z_{\text{hub}} \); and vehicle repositioning costs: \( z_{\text{reposition}} \). Formulae for these components are developed in Sections 3.1 - 3.5. The sum (integral) of \( z \) across all items over the planning horizon, given in Section 3.6, is an approximation of the total system cost.

The following cost constants are used\(^8\):

- \( c_d^u \) cost of overcoming distance, for vehicle of type \( u \in V \) (\$/distance)

- \( c_d'^u \) marginal transportation cost per item, for vehicle of type \( u \in V \) (\$/item * trip)

\(^8\)For ease of illustration, facilities are shown to have the same costs; costs do vary by facility type in Section 5.
cost of stopping a vehicle of type \( u \in V \) at a terminal or customer ($/stop$)

\( c_f \) annualized fixed terminal cost ($/terminal$)

\( c_f \) terminal handling cost per item ($/item$)

\( c_h \) storage (rent) cost for items ($/item*time$)

In addition, the following auxiliary functions are used:

\( \lambda_s^i(x) \) trip attraction rate (inbound flow) in a region of unit area about \( x \) (items/unit area*time);

\[ \lambda_s^i(x) = \int_{x^o \in A} \lambda^s(x^o, x) dx^o, \quad \text{for } s \in S \]

\( \lambda_s^o(x) \) trip generation rate (outbound flow) about \( x \) (items/unit area*time);

\[ \lambda_s^o(x) = \int_{x^i \in A} \lambda^s(x, x^i) dx^i, \quad \text{for } s \in S \]

\( \lambda_b^A(x) \) directional air network demand, \( \lambda_b^A(x) = \lambda_b^E(x) + \omega_b(x) \lambda_b^D(x) \), for \( b = i, o \)

\( \lambda_b^G(x) \) directional ground network demand, \( \lambda_b^G(x) = (1 - \omega_b(x)) \lambda_b^D(x) \), for \( b = i, o \)

\( \lambda_m^T(x) \) bidirectional network-specific demand, \( \lambda_m^T(x) = \sum_{b \in B} \lambda_b^m(x) \), for \( m = A, G \)

\( \lambda_b(x) \) directional demand for combined networks, \( \lambda_b(x) = \sum_{m=A,G} \lambda_b^m(x) \), for \( b \in B \)

\( \lambda_T(x) \) bidirectional demand for combined networks, \( \lambda_T(x) = \sum_{b \in B} \lambda_b(x) \)

\( \delta(x) \) total customer density for combined networks, \( \delta(x) = \sum_{s \in S} \delta^s(x) \)

### 3.1 Local transportation costs

Local costs account for pickup and delivery costs between origins/destinations and consolidation terminals. In the morning, delivery vehicles depart from a consolidation terminal and complete their deliveries. Vehicles that will be used for pickup tours in the afternoon are then repositioned without returning to the terminal, and the rest return. In the afternoon, pickup tours are conducted and
then vehicles return to the consolidation terminal. It is assumed that pickup and delivery routes are designed independently. Repositioning costs are covered in Section 3.4. Figure 4 illustrates the local distribution process for (a) semi-integrated networks and (b) fully integrated networks.

![Figure 4: Local distribution](image)

We introduce a function $f(r, v, n, \delta)$ to designate the average unit cost for items delivered in batches of size $v$ on VRP routes making $n$ stops to customers of density $\delta$ and $r$ distance units away from a depot. It is known (see, e.g., Daganzo (1999)) that:

$$f(r, v, n, \delta) \approx c_d + \frac{r c_d + c_q}{n v} + \left(\frac{n - 1}{n}\right) \left(\frac{c_d k(\delta)^{-\frac{1}{2}} + c_q}{v}\right)$$  \hspace{1cm} (4)$$

where $k$ is a constant dependent on the distance metric; $k \approx 0.8$ for grids\textsuperscript{9}. The first component of (4) represents the cost of moving items in and out of the vehicle. The second term represents the linehaul cost of travel from the depot to the customer region with a stop at the depot. The last term represents the local detour cost of travel between customers in the region, including stops.

The cost expression and constraints for each routing direction $b$ and network type $m$ are then:

$$z_{\text{local}}^{m,b}(x) = f(r_0^{m}(x), v_0^{m,b}(x), n_0^{m,b}(x), \delta^{m}(x))$$  \hspace{1cm} (5a)

\textsuperscript{9}The parameter $k$ can be estimated through simulation, see Daganzo (1999).
Equation (5b) ensures that loads do not exceed vehicle capacity. Equation (5c) ensures that routes have at least one stop, and prohibits long routes. Its upper bound is used instead of a time constraint on the length of a shift to avoid the introduction of more notation.\textsuperscript{10} Equation (5d) ensures that customers are visited with a minimum frequency. Equation (5e) expresses the shipment size $v_{0}^{m,b}(x)$ as a function of the demand accumulated during a headway using Little’s formula ($\delta^* (x) = \delta (x)$ for integrated networks). Equations (5d) and (5e), combined, limit the shipment sizes. Equation (5f) expresses the dependence between linehaul distance and terminal density as stated in equation (2).

Decision variables $v_{0}^{m,b}(x)$ and $r_{0}^{m}(x)$ are uniquely determined by $h_{0}^{m,b}(x)$ and $\Delta C(x)$, respectively, and can be removed from the formulation. However, the presentation is cleaner if they are retained until Section 4.

With non-integrated and semi-integrated networks, four copies of $z_{local}^{m,b}(x)$ appear in expression (3). With fully integrated networks, only two copies of $z_{local}^{b}(x)$ (one for each direction) appear. The same is true of constraints.

\textsuperscript{10}Since it is often the number of stops that limits the number of items in a vehicle, a rough approximation of the average batch size is sufficient.
3.2 Access transportation costs

Access tours between consolidation terminals and breakbulk terminals or airports are similar to local tours. Again, the VRP approximation is used. The average access cost per item is:

\[
z_{\text{access}}^m(x) = f(r_1^m(x), v_1^m(x), n_1^m(x), \Delta C(x))
\] (6a)

subject to:

\[n_1^m(x)v_1^m(x) \leq V_1
\] (6b)

\[1 \leq n_1^m(x) \leq N^m_1
\] (6c)

\[h_1^m(x) \leq H_1^m
\] (6d)

\[v_1^m(x) = \frac{\lambda^m_b(x)}{\Delta C(x)} h_1^m(x)
\] (6e)

\[r_1^G(x) = \frac{2}{3}(\pi \Delta_B(x))^{-\frac{1}{3}} \quad r_1^A(x) = \frac{2}{3}(\pi \Delta_P(x))^{-\frac{1}{3}}
\] (6f)

\[r_1^m(x), v_1^m(x), h_1^m(x) > 0
\] (6g)

For fully integrated networks, one must specify the network demand rates that appear in (6e) with another equation since the network demand rates no longer equal the service level demand rates. Recall that \(\lambda^m_b(x) = \lambda^F_b(x) + \omega_b(x)\lambda^D_b(x)\) and \(\lambda^G_b(x) = (1 - \omega_b(x))\lambda^D_b(x)\) for \(b \in B\), where \(\omega_b(x) \geq 0\). Excess aircraft capacity determines the values of \(\omega_b(x)\) and this is discussed next.

3.3 Long haul transportation costs

3.3.1 Air network

Since all items traveling by air are served through one main hub, the problem decomposes into a many-to-one distribution problem inbound to the hub, and a similar one-to-many problem outbound. In both directions, we use the VRP approximation of one depot (the air hub) serving several customers (the airports). Operating headways are restricted to one day \((h_2^A = \bar{h} = 1\) day),
and are not decision variables. The average linehaul distance, \( r_2^A(x) \) is simply the distance from \( x \) to the hub which depends on the location of the main hub. As shown in Smilowitz (2001), it may be inefficient to operate a symmetric air network (inbound trips to a region mirror outbound trips from that region). Thus, inbound and outbound long haul trips are modeled separately.

\[
z_{longhaul}^{A,b}(x) = f(r_2^A(x), v_2^{A,b}(x), n_2^{A,b}(x), \Delta_P(x))
\]

subject to:

\[
n_2^{A,b}(x)v_2^{A,b}(x) \leq V_2^A \quad (7b)
\]

\[
1 \leq n_2^{A,b}(x) \leq N_2^A \quad (7c)
\]

\[
v_2^{A,b}(x) = \frac{\lambda_0^A(x)}{\Delta_P(x)} \tilde{h} \quad (7d)
\]

\[
v_2^{A,b}(x) > 0 \quad (7e)
\]

\[
\Delta_P(x) \geq \frac{1}{\rho^2 \pi} \quad (7f)
\]

Constraint (7f) ensures that the service radius from an airport does not exceed a maximum distance \( \rho \) to guarantee the timely completion of access and local tours.

For semi-integrated networks, expressions (7) are used only for express items, \( \lambda_0^A(x) \equiv \lambda_0^E(x) \). With integrated routing, a fraction of deferred items may travel by air, provided excess capacity exists. Constraint (7g) is added to restrict the amount shifted \( \omega_b(x) \) by the available capacity.

\[
\frac{n_2^{A,b}(x)\tilde{h}}{\Delta_P(x)} \left( \nu \omega_b(x) \lambda_0^D(x) + \lambda_0^E(x) \right) \leq V_2^A, \quad \text{for } \nu \geq 1, \omega_b(x) \leq 1 \quad (7g)
\]

The constant \( \nu \) is added because it is not economical to fill aircraft with deferred items to the same capacity level as with more profitable express items since operating costs increase with load size.

3.3.2 Ground network

The ground network contains multiple breakbulk terminals; therefore, the problem cannot be decomposed in the same manner. Fortunately, continuum approximation models for many-to-many
non-integrated systems with breakbulk terminals have been developed. It has been shown (see Daganzo (1999)) that the vehicle distance traveled between breakbulk terminals can be easily estimated, without specifying the exact routing of items, when it can be assumed that vehicles travel full. Further, as the number of terminals in the service region \( \int_{x \in A} \Delta B(x) dx \) increases, the linehaul component of this distance rapidly approaches the ratio of the total item-miles demanded and the vehicle capacity. The linehaul component is thus independent of all decision variables. It is not included in the cost model, but added as a constant to the final cost. Vehicles can travel full either by increasing headways or visiting multiple terminals \( n^G_2(x) v^G_2(x) = V^G_2 \). The detour component associated with multiple stops is estimated by:

\[
\left( \frac{n^G_2(x) - 1}{n^G_2(x)} \right) \left( \frac{c_t k(\Delta B(x))^{-\frac{1}{2}} + c_q}{v^G_2(x)} \right)
\]  

(8a)

Shipment size is approximated as follows. The average ground network demand rate for the entire service region across all origins and destinations is \( \bar{\lambda}^G = \int_{x^o \in A} \int_{x^i \in A} \frac{\lambda^G(x^o, x^i)}{|A|^2} dx^i dx^o \) where \( \lambda^G(x, x^i) \) is the deferred demand rate \( \lambda^D(x, x^i) \) reduced by the total amount shifted to air). The average breakbulk terminal density is \( \bar{\Delta}_B = \int_{x \in A} \frac{\Delta B(x)}{|A|} dx \). Averaged over all destinations, the average shipment size collected from a breakbulk terminal at \( x \) is approximated by

\[
v^G_2(x) \approx \frac{\bar{\lambda}^G |A| h^G_2(x)}{\bar{\Delta}_B(x)}.
\]

(8b)

### 3.4 Vehicle repositioning costs

#### 3.4.1 Local and access levels

On local tours, the number of vehicles dispatched for morning deliveries may be insufficient to cover afternoon pick-up; extra empty vehicles must be deployed for collection. Conversely, vehicles may return empty to the consolidation terminal after morning distribution if inbound demand exceeds outbound demand. The same is true for access trips. It is assumed that all local and access
tours operate from one terminal. Thus, the number of repositioning trips is simply the number of vehicles needed to serve the demand imbalance, $|\lambda^m_{\text{out}}(x) - \lambda^m_{\text{in}}(x)|$. For vehicles with capacity $V$ operating in a region of terminals with density $\Delta(x)$, the number of trips is $\frac{|\lambda^m_{\text{out}}(x) - \lambda^m_{\text{in}}(x)|}{\Delta(x)V}$. The total repositioning cost is therefore equal to the number of trips multiplied by the distance cost and the average distance from the terminal to the customers, $r(x)$. Prorating this cost by the total items served by the terminal, $\frac{\lambda^m_{\text{out}}(x)}{\Delta(x)}$, and replacing $r(x)$ with $\frac{2}{3}(\pi \Delta(x))^{-\frac{1}{2}}$, the average cost per item is:

$$\frac{2}{3} c_d |\lambda^m_{\text{out}}(x) - \lambda^m_{\text{in}}(x)| \frac{\lambda^m_{\text{out}}(x)}{\Delta(x)V} (\pi \Delta(x))^{-\frac{1}{2}}.$$  (9)

### 3.4.2 Long haul level

Repositioning empty vehicles between breakbulk terminals is more difficult to model since demand imbalances between breakbulk terminals require vehicle repositioning between terminals. If demand is balanced, the repositioning term between breakbulk terminals is zero. The added repositioning costs due to systematic imbalances can be bound tightly from above by a function of the origin-destination demand table, independent of all decision variables; see Daganzo and Smilowitz (2003). Therefore, long haul repositioning is not included in the optimization phase, but is added to the total system cost.

### 3.5 Terminal costs

Terminal costs consist of handling costs, facility charges, and storage fees. The value of decision variables and parameters vary across terminal types, but the functional form of the terminal cost is the same. This section introduces the generic cost model; specific costs for each terminal type are included in Section 3.6.

Consider a terminal serving an inbound and outbound flow of $Q(x)$ items per unit time given
by the trip attraction and generation, $Q(x) = \frac{\lambda_i(x)}{\Delta y(x)} + \frac{\lambda_o(x)}{\Delta s(x)}$. The cost per item for a terminal is

$$g(Q(x), h_o(x), h_i(x)) = c_f + \frac{c_f}{Q(x)} + \sum_{b=i,o} c_h h_b(x)$$

(10)

The first term represents handling costs, and the second represents a fixed cost per terminal which is prorated by the flow $Q(x)$. The final term represents a storage cost dependent on the length of time an item is held at a terminal. It is assumed that this length of time is proportional to the routing headways, $h_b(x)$, assuming routes are not coordinated; see Daganzo (1999).

### 3.6 Complete model

A complete logistic cost function, containing all transportation and terminal costs, is used to obtain optimal designs and compare integration strategies. This function integrates the cost components described in the previous sections over the item flow in the service area. The expression below is the result for a fully integrated network. By setting $\omega_b(x) = 0, \forall x \in A$, and $b \in B$ and separating local costs by network, the expression represents a semi-integrated network. The function is:
The integrand of (11) begins with local transportation costs, summing both collection and delivery costs. The next line represents access costs for trips to and from airports and breakbulk terminals. The following two lines represent long haul costs for air and ground transportation, respectively. The next line includes the repositioning costs for local and access vehicles. The final two lines include terminal costs. The goal is to choose the decision functions that minimize (11) subject to constraints defined in the previous subsections. The problem is reduced in Section 4 to a series of subproblems that can be easily programmed into a spreadsheet.
3.7 Stochastic demand

The continuum formulation can be extended easily to account for stochastic demand. Here we highlight how this is done; for further details see Smilowitz (2001). Due to different characteristics, uncertainty is treated differently for each mode and service level. We introduce slack in the capacity constraints for the air network and add additional repositioning costs in the ground network.

Demand variations are modeled as a stationary process with independent increments and a location-dependent index of dispersion (variance-to-mean ratio). More specifically, the demand in any time interval between any two regions of small area (e.g., about points \( x^o \) and \( x^i \)) are assumed to be independent of other demands if at least one of the following conditions is satisfied: (1) the two origin areas do not overlap; (2) the two destination areas do not overlap; (3) the two time intervals do not overlap. Inbound and outbound demands in a region have variance-to-mean ratios \( \gamma^s_b(x), s \in S, b \in B (\text{items}) \). It is assumed that these values do not vary over time.

In the air network, the system is overdesigned to minimize the possibility of express demand exceeding capacity. Thus, in the design process, \( V_2^A \) is replaced with a smaller quantity \( \theta^{A,b} V_2^A \) for some positive \( \theta^{A,b} < 1 \), such that:

\[
\theta^{A,b} V_1^A + 3 \sqrt{\theta^{A,b} \gamma^{E,b} V_2^A} \leq V_2^A
\]

Across many days, this leaves an average excess capacity of \((1 - \theta^{A,b}) V_2^A\) in all aircraft, equivalent to three standard deviations of the expected vehicle load, but ensures that overflows would be unlikely. This slack is added to (7g) to define the expected shift amount of deferred items to air.

In the ground network, additional strategies to handle uncertainty, such as rerouting vehicles, are available due to relaxed time constraints. Vehicles can still travel full, although routing may change slightly from day to day. This does not change the full vehicle miles traveled at all levels, nor does it change the peddling costs. However, the need for empty vehicle repositioning increases since
empty vehicles may be rerouted to accommodate demand fluctuations between terminals.\footnote{No repositioning of delivery vans occurs between consolidation terminals because it is assumed that a sufficient supply of vans exists at each terminal and demand fluctuations can be absorbed by holding items across days.} The vehicle repositioning cost due to stochastic effects alone can be approximated as a transportation problem, as shown in Daganzo and Smilowitz (2003), and added to deterministic repositioning costs.\footnote{This is conservative since this sum is the average cost obtained by a superposition of the deterministic solution and the TLP solution including only the stochastic deviation from the mean, which is a feasible (sub-optimal) solution of the real problem.} The average number of total empty vehicle miles across many days required to reposition trucks at the least cost each day is a function of the total area of the service region $A$, the number of terminals, $\Delta_y(x)|A|$, and $\sigma_y(x)$, where $\sigma_y(x)$ is the standard deviation of the flows inbound to and outbound from a terminal equal to $\sqrt{\gamma^C(x)\lambda^C(x)+\gamma^O(x)\lambda^O(x)}$. The stochastic repositioning cost per terminal per unit time is:

$$z_{\text{reposition}} = c^t_d \sigma_y(x) \Delta_y(x)^{-\frac{1}{2}} \left( 1 + 0.078 \log_2(\Delta_y(x)|A|) \right)$$

(13)

In order to apply the area-decomposition solution technique, the expected terminal density must be replaced with the local terminal density.

4 Solution method

This section describes how the problem can be separated into a series of subproblems that can be solved in closed form. First (11) is simplified by expressing it in terms of only the terminal densities and operating headways. The subproblems are presented in Sections 4.1 - 4.4.

Recall that ground vehicles must either reach their capacity or their maximum number of stops; i.e., either (5b) or (5c) must be binding (on the upper side). The same is true for (6b) and (6c), and for (7b) and (7c). Therefore, we can replace $n^{m,b}_i(x)$ with $\min\left\{ N^m_i, \frac{V^m_i}{\lambda^m_i(x)} \right\}$. Equations
(5d), (6d), and (7d) are used to eliminate $v_{l^m,b}(x)$ and equations (5e) and (6e) to eliminate $r_{l^m}(x)$. Recall that $\omega_b(x)$ can be set equal to the largest possible value consistent with (7g) and eliminated. Therefore, only terminal densities and operating headways remain. The complete model is presented below in a compact form that highlights these decision variables. Expressions for the constants $\alpha$, $\beta$, $\chi$, $\kappa$, and $\Pi$ are given in the appendix. For further economy of notation, the dependence of these constants on $x$ is not explicitly stated. The complete model is:

$$\begin{align*}
\min z &= \int_{x \in A} \left\{ \sum_{b=i,o} \left( \alpha_1^b h_0^b(x) + \alpha_2^b (h_0^b(x))^{-1} \right) + \beta_1 \Delta_C(x)^{-\frac{1}{2}} 
+ \sum_{b=i,o} \sum_{m=A,G} \left( \beta_2 \frac{\Delta_C(x)}{h_{m,b}^1(x)} + \beta_3 \frac{\Delta_C(x)}{h_{m,b}^2(x)} + \beta_4 h_{m,b}^1(x) \right) + \beta_6 \Delta_C(x) 
+ \chi_1 \Delta_P^{-\frac{1}{2}}(x) + \chi_2 \Delta_P(x) + \chi_3 \Delta_P^+(x)
+ \kappa_1 \Delta_B^{-\frac{1}{2}}(x) + \kappa_2 \Delta_B(x) + \kappa_3 \frac{\Delta_B(x)^{\frac{1}{2}}}{h_2^G(x)} + \kappa_4 \frac{\Delta_B(x)^{\frac{1}{2}}}{h_2^G(x)} + \kappa_5 h_2^G(x) + \Pi \right) \right\} dx \quad (14a)
\end{align*}$$

subject to:

$$\begin{align*}
\frac{\lambda_b(x)\delta(x)}{N_0 V_0} &\leq h_0^b(x) \leq \frac{\lambda_b(x)\delta(x)}{V_0} \quad \forall b \in B \quad (14b)
\frac{\lambda_m^b(x)}{V_1} &\leq \frac{\Delta_C(x)}{h_{m,b}^1(x)} \leq \frac{N_1^m \lambda_m^b(x)}{V_1} \quad \forall b \in B; m = A, G \quad (14c)
\frac{\lambda_A^b(x)}{V_2^A} &\leq \frac{\Delta_P(x)}{h_{m,b}^1(x)} \leq \frac{N_2^A \lambda_A^b(x)}{V_2^A} \quad \forall b \in B \quad (14d)
\frac{\chi^G |A|}{V_2^G} &\leq \frac{\Delta_B(x)}{h_2^G(x)} \leq \frac{N_2^G \chi^G |A|}{V_2^G} \quad (14e)
0 &< h_{l^m,b}^1(x) \leq H_{l^m} \quad \forall b \in B; m = A, G; l \in L \quad (14f)
\Delta_P(x) &\geq \frac{1}{\rho^2 \pi} \quad (14g)
\end{align*}$$

Note that (14) can be decomposed into five classes of subproblems that involve the following groups
of decision functions:

1\textsuperscript{o} local outbound headways, $h^0_b(x)$

1\textsuperscript{i} local inbound headways, $h^i_b(x)$

2 consolidation terminal densities and access headways, $\Delta_C(x), h^{m,b}_1(x)$

3 airport densities, $\Delta_P(x)$

4 breakbulk terminal densities and long haul ground headways, $\Delta_B(x), h^G_2(x)$.

The subproblems in each class are analyzed below.

4.1 Subproblems 1\textsuperscript{o} and 1\textsuperscript{i}: local headways

For $b = i, o$, these subproblems are:

\[
\min z_{1b} = \int_{x \in A} \left( \alpha_1^b h^b_0(x) + \frac{\alpha_2}{h^b_0(x)} \right) dx
\] (15a)

subject to:

\[
\frac{\lambda_b(x)\delta(x)}{N_0V_0} \leq h^b_0(x) \leq \frac{\lambda_b(x)\delta(x)}{V_0}
\] (15b)

\[0 < h^b_0(x) \leq H_0
\] (15c)

Note that (15) can be further decomposed by $x$ because the integrand and the constraints are local in nature. Hence, one can simply minimize the integrand of (15a) for every $x$ and sum across all such subdivisions of the total area. This is the geographic decomposition mentioned earlier. Since the integrand is a simple economic order quantity (EOQ) problem, the optimal headway can be expressed in a simple form. It is $h^b_0(x)^* = \sqrt{\frac{\alpha_2}{\alpha_1}}$ if this value satisfies the constraints. Otherwise, it is one of the extreme points defined by the constraints. The result should be intuitive. With higher transportation costs, headways should be lengthened, and with higher rent costs, headways
should be shortened. We find that for reasonable values of the parameters, \((15c)\) is binding at its upper bound.

4.2 Subproblem 2: consolidation terminal densities and access headways

Subproblem 2 is:

\[
\min z_2 = \int_{x \in A} \left( \beta_1 \Delta_C(x)^{-\frac{1}{2}} + \sum_{b=i, o} \sum_{m=A, G} \left( \beta_2 \frac{\Delta_C(x)}{h_1^{m,b}(x)} + \beta_3 \frac{\Delta_C(x)}{h_1^{m,b}(x)} + \beta_4 h_1^{m,b}(x) \right) + \beta_6 \Delta_C(x) \right) dx \quad (16a)
\]

subject to:

\[
\frac{\lambda_b^m(x)}{V_1} \leq \frac{\Delta_C(x)}{h_1^{m,b}(x)} \leq \frac{N_1^m \lambda_b^m(x)}{V_1} \quad b \in B; m = A, G \quad (16b)
\]

\[
0 < h_1^{m,b}(x) \leq H_1^m \quad b \in B; m = A, G \quad (16c)
\]

On the surface, subproblem 2 is more complicated than subproblem 1 because it involves a non-convex objective function and non-linear constraints. Fortunately, the following changes of variable transform \((16)\) into a convex problem with linear constraints: \(w_C(x) = \ln(\Delta_C(x))\), \(w_{m,b}(x) = \ln(h_1^{m,b}(x))\), \(b \in B; m = A, G\). The transformed problem is:

\[
\min z_2' = \int_{x \in A} \left( \beta_1 e^{-\frac{w_C(x)}{2}} + \sum_{b=i, o} \sum_{m=A, G} \left( \beta_2 e^{\frac{w_C(x)}{2} - w_{m,b}(x)} + \beta_3 e^{w_C(x) - w_{m,b}(x)} + \beta_4 e^{w_{m,b}(x)} \right) + \beta_6 e^{w_C(x)} \right) dx \quad (17a)
\]

subject to:

\[
\ln \left( \frac{\lambda_b^m(x)}{V_1} \right) \leq w_C(x) - w_{m,b}(x) \leq \ln \left( \frac{N_1^m \lambda_b^m(x)}{V_1} \right) \quad b \in B; m = A, G \quad (17b)
\]

\[
w_{m,b}(x) \leq \ln(H_1^m) \quad b \in B; m = A, G \quad (17c)
\]

This subproblem can also be decomposed by location. Since the transformed problem is convex, it can be solved with gradient search techniques.
The same spatial decomposition and logarithmic transformation techniques reduce subproblems 3 and 4 to simple convex programs.

4.3 Subproblem 3: airport densities

The following change of variable is introduced: \( w_P(x) = \ln(\Delta_P(x)) \). Subproblem 3 is then:

\[
\min z_3^0 = \int_{x \in A} \left( \chi_1 e^{-\frac{w_P(x)}{2}} + \chi_2 e^{w_P(x)} + \chi_3 e^{-\frac{w_P(x)}{2}} \right) dx
\]

subject to:

\[
\ln \left( \frac{\lambda_0^A(x)}{V_2^A} \right) \leq w_P(x) \leq \frac{\ln \left( \frac{N_2^A \lambda_0^A(x)}{V_2^A} \right)}{b \in B} \tag{18b}
\]

\[
w_2(x) \geq \ln \left( \frac{1}{\rho^2 \pi} \right) \tag{18c}
\]

This can be decomposed geographically by \( x \) and easily solved.

4.4 Subproblem 4: breakbulk terminal densities and long haul ground headways

The terminal density and headway variables are transformed as follows: \( w_B(x) = \ln(\Delta_B(x)), \) \( w_2(x) = \ln(h_2^G(x)) \). This results in:

\[
\min z_4^0 = \int_{x \in A} \left( \kappa_1 e^{-\frac{w_B(x)}{2}} + \kappa_2 e^{w_B} + \kappa_3 e^{-\frac{w_B(x)}{2}} - w_2(x) + \kappa_4 e^{w_B(x) - w_2(x)} + \kappa_5 e^{w_2} \right) dx \tag{19a}
\]

subject to:

\[
\ln \left( \frac{\lambda_0^G |A|}{V_2^G} \right) \leq w_B(x) - w_2(x) \leq \frac{\ln \left( \frac{N_2 \lambda_0^G |A|}{V_2^G} \right)}{b \in B} \tag{19b}
\]

\[
w_2(x) \leq \ln(H_2^G) \tag{19c}
\]

This subproblem can again be decomposed by location and solved easily. Hence, the entire problem has been reduced to a series of easily solved convex subproblems.
4.5 Model accuracy

Unfortunately the accuracy of the complete design methodology cannot be accurately ascertained without the ability to solve the discrete problem; see Section 2; however component cost models presented here are accurate enough for design purposes. The literature also contains other validation tests. Vehicle routing cost formulae have been validated in Robuste et al. (1990) and Erera (2000). They show that the costs of advanced local distribution strategies can be approximated well within 5% of those arising from discrete simulations for problems with deterministic and stochastic demand. Costs from hub location decisions obtained with continuum approximations have also been validated against discrete cost models in the literature; see Campbell (1993). Long haul operating costs are validated using a discrete formulation from Smilowitz et al. (2003) given a fixed network of terminals. The difficulties encountered when solving large problem instances highlight the complexity of formulation (1). Empty vehicle repositioning costs are validated with TSP simulations in Daganzo and Smilowitz (2003). Terminal cost approximations are validated in Smilowitz (2001). In all these cases, errors are substantially less than 5%; often less than 1%. Therefore, we can conclude that the difference between the optimum cost predicted by our model and the true optimum cannot exceed 5% and it is likely much smaller, see proof in Appendix B.

5 Case study

The design methodology introduced in Sections 3 and 4 has been employed in Smilowitz (2001) to obtain network configurations for large package delivery networks roughly the size of the contiguous United States. Operating costs and statistics are derived from Kiesling (1995) and company literature. Population density is used as a proxy for package demand rates ($\lambda$) and housing density as a proxy for customer density ($\delta$). The 1990 U.S. census includes population, housing counts,
land area and geographic coordinates for all Metropolitan Statistical Areas (MSA) in the United States; see U.S. Census Bureau (1990). The largest MSA’s are aggregated into groups of common demand and geographic features to form twenty subregions. The subregions are large enough to contain multiple terminals, yet small enough such that average network characteristics (demand levels, distances to main air hub, etc.) are representative of the entire subregion. The entire service region is 2,500,000 \textit{mi}^2. The areas of the twenty subregions range from 16,383 \textit{mi}^2 to 225,974 \textit{mi}^2.

We focus on the following question: given a pair of non-integrated networks for deferred and express demand, when does integration significantly reduce costs? In a series of test cases, deferred demand levels are increased with the goal of determining the level of deferred demand required to justify integration. Both deterministic and random demands are considered. In Figure 5, the total savings achieved through integration is plotted as a percent of the total pre-integration cost of the original air network. The x-axis represents the average daily deferred demand and the figure indicates the point at which deferred demand exceeds express demand. An average demand of 1.8 million packages per day are assumed for express items. Deferred demand ranges from 0 (no integration) to 5.2 million packages per day. On the y-axis, the total (air and ground) network cost savings divided by the total pre-integration air network costs are plotted.

As the figure indicates, the benefits of integration increase as deferred demand increases. Express carriers may be reluctant to integrate operations with deferred carriers when the level of deferred demand is significantly less than that of express. However, savings grow quickly as deferred demand increases, even before deferred demand equals express demand. The growth rate of savings decreases and savings reach an asymptote once excess air capacity is filled and the maximum benefits of local transportation integration are realized. In addition, the figure reveals that savings increase significantly when demands are uncertain. Additional savings may be achieved when it is necessary to overcapacitate the air network for seasonal demand fluctuations. This insight helps to explain
the different business strategies of United Parcel Services (UPS) and Federal Express. UPS has adopted a more integrated strategy than Federal Express. A large deferred carrier such as UPS should realize greater cost savings from integration.

The test cases clearly indicate the dominance of local transportation costs. This is not surprising since local transportation consists of many trips made in small vehicles operating on short headways. In turn, changes in local costs have a large impact on total cost. As expected, large savings in local transportation costs are realized with integrated routing as a result of higher customer density, see Figure 4(b) versus Figure 4(a). The total savings are greatest in regions of the service network where local transportation costs account for over 45% of total ground and air network operating costs. These regions typically have low demand levels and low customer densities. With the rise of e-commerce, the importance of local distribution to individual customers should increase and the incentive for integrating local distribution between service levels should increase too.

Additional analysis of test cases suggests that merging infrastructure from existing networks to form an integrated network yields cost savings comparable with designing an entirely new integrated network. Across all test cases, the largest difference in total cost between integration strategies
with existing infrastructure and redesigning a network is only 0.5% which hardly justifies the cost of relocating or building facilities.

Of course, there are other costs and benefits to integration not considered here that could impact decisions. Integration gives carriers the ability to move deferred items quickly in response to routing problems in the ground network (weather, surge in demand, etc.). Further, overhead costs including administrative costs, sales costs, etc. can be reduced through integration when a delivery firm can use one office to multiple service levels. However, there may be additional costs of complexity involved with integration.

6 Conclusions and future work

The proposed methodology can be used to design complex integrated distribution systems with multiple service levels and multiple transportation modes. We show that the problem can be reduced to a series of simple subproblems while considering all key costs (facility charges and vehicle repositioning, as well as transportation and inventory) for distribution that includes multiple transshipments, peddling tours, and shipment choice. We have estimated errors from approximations to be within 5%. Importantly, although we cannot solve the problem exactly, we can use the approximation models to make key strategic decisions.

This research addresses significant gaps identified in the continuum approximation literature. This is helpful since continuum approximation modeling can be a powerful tool used in conjunction with discrete optimization. The results of continuum approximation can be used to establish guidelines for network design and routing, as well as to evaluate the merits of various integration strategies. By first approximating cost savings from integration, one can decide if the magnitude of the savings warrants more detailed discrete optimization. In addition, results from continuum approximation can give insights for discrete optimization.
The solution to the continuous optimization problem (COP) provides sufficient detail to determine fleet size and other discrete decision variables. The network design is obtained by partitioning the service region into “round” service regions of approximate size $\Delta_y(x)^{-1}$ and locating terminals at their centroids. For example, consider one subregion of approximately 16,000 square miles (roughly the size of Northern Illinois). Under moderately dense demand assumptions (two customers per square mile requesting two items per customer), the COP results require sixteen consolidation terminals and one breakbulk terminal to serve the region. Local distribution tours should make 22 stops per tour on average. This translates to approximately 76 delivery vehicles assigned to each consolidation terminal. Each consolidation terminal should deliver three truckloads of items to the breakbulk terminal. The solution also reveals that the average consolidation terminal should be 260 miles from the breakbulk terminal. Assuming a speed of 50 miles per hour, a vehicle could perform at most two round trip access tours per day. Therefore, three vehicles are required to serve two consolidation terminals and a total of 24 vehicles would be need to serve the 16 consolidation terminals. Similar calculations can be performed for other regions resulting in design and operational guidelines for the complete service area. These guidelines can be then be used to geographically separate discrete terminal location problems and vehicle and item routing problems, as shown in Smilowitz (2001). Ouyang and Daganzo (2003) develop algorithms to translate continuous approximation outputs into discrete solutions.

Acknowledgments

The authors would like to thank Anslem Brecht at the University of Frankfurt for valuable input on the longhaul cost models.
References


Geoffrion, A. (1976). The purpose of mathematical programming is insight, not numbers. Interfaces, 7(1), 81–92.


Appendix A: Notation

Discrete variable arrays

\[
Y^1 = (Y^1_1, Y^1_2, ..., Y^1_I)
\]

\[
Y^2 = (Y^2_1, Y^2_2, ..., Y^2_I)
\]

\[
X^0 = (X^0_1, X^0_2, ..., X^0_I)
\]

\[
X^1 = (X^1_1, X^1_2, ..., X^1_I)
\]

\[
X^2 = (X^2_1, X^2_2, ..., X^2_I)
\]

\[
Z^0 = \begin{pmatrix}
Z^0_{1,1} & Z^0_{1,2} & \cdots & Z^0_{1,I_1} \\
Z^0_{2,1} & Z^0_{2,2} & \cdots & Z^0_{2,I_1} \\
\vdots & \vdots & \ddots & \vdots \\
Z^0_{I_0,1} & Z^0_{I_0,2} & \cdots & Z^0_{I_0,I_1}
\end{pmatrix}
\]

\[
Z^1 = \begin{pmatrix}
Z^1_{1,1} & Z^1_{1,2} & \cdots & Z^1_{1,I_2} \\
Z^1_{2,1} & Z^1_{2,2} & \cdots & Z^1_{2,I_2} \\
\vdots & \vdots & \ddots & \vdots \\
Z^1_{I_1,1} & Z^1_{I_1,2} & \cdots & Z^1_{I_1,I_2}
\end{pmatrix}
\]

Network sets

\(S\) Set of service levels, \(S = \{E, D\}\) for express and deferred items.

\(L\) Set of distribution levels, \(L = \{0, 1, 2\}\): local (0), access (1) and long haul (2).

\(B\) Set of route directions, \(B = \{i, o\}\) for trips inbound to and outbound from a terminal.

\(V\) Set of vehicle types, for simplicity \(V = \{a, t\}\), for air and truck.

\(T\) Set of terminal (node) types, \(T = \{C, B, P, H\}\) for consolidation terminals (C), breakbulk terminals (B), airports (P), and main air hub (H).
Demand Parameters

\( \delta^s(x) \) spatial customer densities for service level \( s \in \mathcal{S} \) (customers/unit area)

\( \lambda^s(x^o, x^i) \) temporal demand rate from a region of unit area about \( x^o \) to a region of unit area about \( x^i \) for service level \( s \in \mathcal{S} \) (items/area\(^2\times\)time)

\( \lambda^t_i(x) \) trip attraction rate in a region of unit area about \( x \) (items/unit area\(^\times\)time); \( \lambda^t_i(x) = \int_{x \in A} \lambda^s(x, x^i) \, dx \)

\( \lambda^g_s(x) \) trip generation rate about \( x \) (items/unit area\(^\times\)time); \( \lambda^g_s(x) = \int_{x \in A} \lambda^s(x^o, x) \, dx \)

Level of Service Parameters

\( H_l \) maximum headway length for a route of type \( l \in \mathcal{L} \) (time)

\( N_l \) maximum number of stops on a route of type \( l \in \mathcal{L} \)

\( V^u \) vehicle capacity for vehicle of type \( u \in \mathcal{V} \) (items)

\( \rho \) maximum airport service radius (distance)

Cost Parameters

\( c^u_d \) costs of overcoming distance, for vehicle of type \( u \in \mathcal{V} \) ($/distance$)

\( c^u_d' \) marginal transportation cost per item, for vehicle of type \( u \in \mathcal{V} \) ($/item\times trip$)

\( c_u^s \) cost of stopping a vehicle of type \( u \in \mathcal{V} \) at a terminal or customer ($/stop$)

\( c_f \) annualized fixed terminal cost ($/terminal$)

\( c_f' \) terminal handling cost per item ($/item$)

\( c_h \) storage (rent) cost for items ($/item\times time$)
Decision functions

\( \Delta_y(x) \) density of terminals of type \( y \in T \) (terminals/unit area)

\( h_{l}^{m,b}(x) \) headway of a route of type \( l \in L \) for network \( m = A, G \) in direction \( b \in B \) (time)

\( n_{l}^{m,b}(x) \) number of stops on a route of type \( l \in L \) for network \( m = A, G \) in direction \( b \in B \)

\( v_{l}^{m,b}(x) \) shipment size per terminal on a route of type \( l \in L \) for network \( m = A, G \) in direction \( b \in B \) (items/terminal)

\( r_{l}^{m}(x) \) average linehaul distance on a route of type \( l \in L \) for network \( m = A, G \) (distance)

\( \omega_b(x) \) fraction of deferred items sent by air for long haul transportation in direction \( b \in B \)

Coefficients and constants

Constant \( \Pi \),

\[
\Pi = \lambda_T(x) \left( c_d^{l} \frac{c_d^{l}k(\delta(x))^{-\frac{1}{2}}}{V_0} + \lambda_T(x)c_d^{l} + \lambda_T(x)c_d^{l} \left( c_d^{l} + 2c_h \hat{h} \right) + \lambda_o(x)c_k \log(2) + 2\lambda_T(x)c_f^{l} \right)
\]

where it is understood that \( \Pi \) is a function of \( x \). Coefficients for local operating headways:

\[
\alpha_1^b = \lambda_b c_h; \ b = i, o \\
\alpha_2 = c_d^{l}k(\delta)^{\frac{1}{2}} + c_d^{l} \delta
\]

Coefficients for consolidation terminal densities and access operating headways:

\[
\beta_1 = \lambda_T \left( \frac{2c_d^{l}}{\sqrt{\pi V_0}} - \frac{c_d^{l}k}{V_1} \right) + \frac{2|\lambda_o - \lambda_i|}{\sqrt{\pi V_0}} c_d
\]

\[
\beta_2 = c_d^{l}k \\
\beta_3 = c_q \\
\beta_4^b = \lambda_b^c c_h \frac{1}{2} \\
\beta_5^b = \lambda_b^G c_h; \ b = i, o \\
\beta_6 = c_f
\]

Coefficients for airport densities:

\[
\chi_1 = \lambda_T \left( \frac{2c_d^{l}}{\sqrt{\pi V_1}} \right) + \frac{2|\lambda_o^A - \lambda_i^A|}{\sqrt{\pi V_1}} c_d \\
\chi_2 = c_f + \sum_{b \in B} \frac{r_2^A c_d + n_2^{A,b} c_q}{n_2^{A,b}} \\
\chi_3 = \sum_{b \in B} \frac{n_2^{A,b} - 1}{n_2^{A,b}} c_d k
\]
Coefficients for breakbulk terminal densities and long haul operating headways:

\[
\kappa_1 = \lambda_o G \left( \frac{2A_c d}{\sqrt{A} V_1} \right) + \frac{2}{\sqrt{2\pi}} \lambda_o G \frac{G_o - G_i}{\sqrt{A} V_1} \left( c_d - \frac{k}{V^2} \right)
\]

\[
\kappa_2 = c_f \quad \kappa_3 = c_d k \quad \kappa_4 = c_q \quad \kappa_5 = \lambda_o G c_h
\]

Appendix B: Proof of cost errors

**Proposition:** If errors in the cost components are less than 5% then the difference between the optimum cost predicted by our model and the true optimum cannot exceed 5%.

**Proof:** Assume \( A, B, C \geq 0 \) and \( p \approx 5\% \); \( |A - \bar{A}| \leq Ap; \ |B - \bar{B}| \leq Bp; \) and \( |C - \bar{C}| \leq Cp \). If \( T = A + B + C \) and \( \bar{T} = \bar{A} + \bar{B} + \bar{C} \) then \( T - \bar{T} = (A - \bar{A}) + (B - \bar{B}) + (C - \bar{C}) \). Therefore \( |T - \bar{T}| \leq |A - \bar{A}| + |B - \bar{B}| + |C - \bar{C}| \leq Ap + Bp + Cp \) (since \( A, B, C \geq 0 \)).

Hence, \( \frac{|T - \bar{T}|}{(A + B + C)} \leq p \)