Cost Modeling and Design Techniques for Integrated Package Distribution Systems

Karen R. Smilowitz*and Carlos F. Daganzo†

December 23, 2005

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.

Key words: Network design problem, continuum approximation, transportation

*Department of Industrial Engineering and Management Sciences, Northwestern University

†Department of Civil and Environmental Engineering, University of California, Berkeley

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 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 across 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 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 between deferred and express package distribution carriers.

Section 1 discusses network configurations and routing principles for package delivery. Section 2 introduces the notation and assumptions of the continuum 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 based on idealizations of the complex 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.[1] Deferred items may be sent by ground or air.

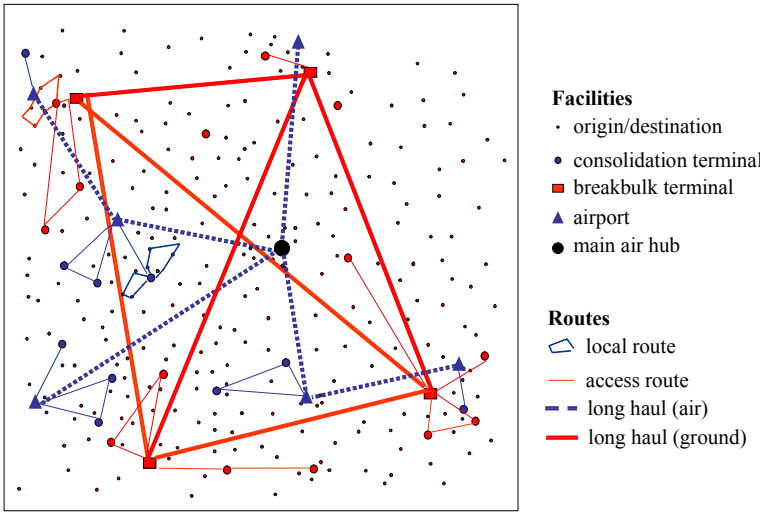

Figure 1: Integrated package distribution network

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

[1]Express items with nearby destinations may not travel by air. Such items are ignored in this study.

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.

The network design problem determines ground and air network configurations (number and location of terminals) and routing guidelines for items and vehicles. We study two distribution strategies.

Fully integrated networks: integrated facilities and routing

Express items are sent by air; deferred items travel either way. Economies of density can yield savings in local transportation. Flexible routing allows excess air capacity to be filled with deferred items, reducing ground transportation needs. Consolidation terminals serve both service types.

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 [1, 3, 25]. As discussed in these and other more general references; e.g., [26], optimal solutions can be found numerically for small instances. In some special cases, it is possible to solve large instances. In general, however, as the network size increases, heuristic solution 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 [7], as well as [2, 4, 29]. Yet, collecting demand and cost data for strategic problems 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 [12]. 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 [10]. Early work on approximation methods ([15, 17, 27]) found that approximations provide near optimal solutions and offer valuable insight into operating strategies and network design. Whereas numerical optimization models perform better on smaller instances, the opposite is true with continuum approximations. The larger the instance, the more accurate the approximations become; see [5, 10, 14]. Often continuum formulations 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 trans-

portation networks, the topic has not been explored thoroughly. This paper fills the following methodological gaps identified in [24]: (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 [19, 22]; and (iv) current models do not consider multiple service levels although several studies have considered distribution of time sensitive items, see [8, 9, 20, 23]. Multiple transportation modes have been included in a limited number of models, see [18]. 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 developed with discrete optimization.

2 Continuum approximations for the 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 [11], the performance of the overall distribution system can be improved if the network is designed with the subsequent operating plans in mind. Detailed demand data are rarely available when the planning horizon is long; forecasts typically predict continuous demand rates for broad geographic areas. This leaves two options: (i) discretize the data and run a precise but hard discrete optimization; or (ii) run an easy but approximate continuum optimization and

discretize the solution. A discrete formulation of the network design problem is presented in [32]. While it is possible to quantify facility costs, it is considerably more difficult to quantify transportation costs and operating constraints in a discrete formulation. The expressions depend on terminal locations and demand allocations, and must account for the different ways in which vehicles may be routed. Moreover, even if one is successful in this endeavor, heuristics are needed to solve realistic instances with hundreds or thousands of possible physical locations. The need to model integrated networks complicates the problem even further. Fortunately many of these difficulties can be overcome with a continuum approximation, since, as shown later, continuity allows one to decompose the problem geographically.

In what follows, we develop approximations to solve the strategic network design problem, which incorporate operating strategies. Capturing all complexities of the operating plans is not feasible; however, developing a model that considers even rudimentary operating issues can improve overall system performance. Therefore, we focus on a simplified network design problem, which can be modified to account for specific operating plans, as done in [31] and discussed briefly in Section 6.

In the continuum formulation, the structure of the distribution network over a service area $\mathcal{A}$ is defined by the network topology, and by demand, level of service, and cost data functions (parameters) that may vary with the coordinates, x, of points on the plane. The solution is described in terms of decision functions of location (variables). A summary of the notation is provided in the appendix.

2.1 Network representation

A typical path of an item from origin to destination is shown in Figure 2(a). 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).

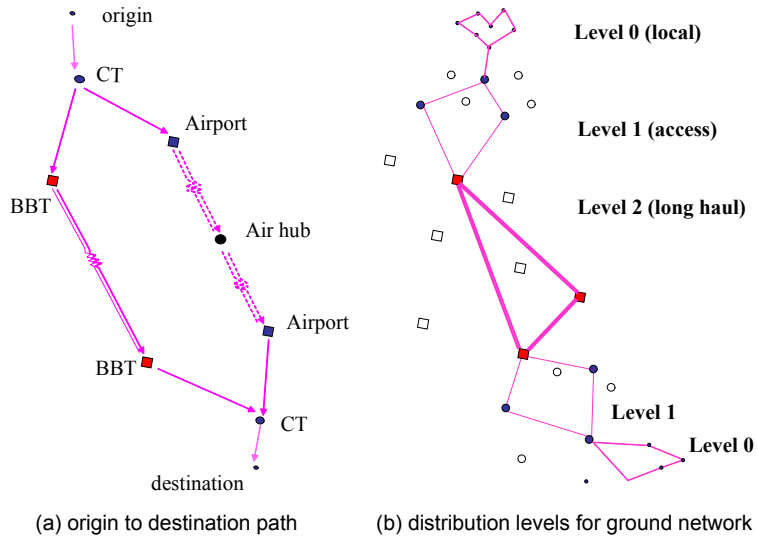

Figure 2: Distribution from origin to destination

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.

2.2 Demand parameters

As discussed earlier, data are often available only on an aggregate level for strategic design problems. Aggregated data can be used to obtain customer density and demand estimates, as shown in [6].

Consistent with the continuum approximation literature on time-sensitive delivery ([8, 9, 20, 23]), we consider stationary, deterministic demand data. Extensions for stochastic demand are shown in Section 3.7. Extensions for seasonal fluctuations such as holiday demand surges are presented in [31]. Systematic demand variations over the course of the day can also be incorporated. However, since pickup and delivery operations are concentrated over a few hours and the system operates on a daily cycle, the demand distribution specific by time-of-day should not significantly influence operating costs, conditional on total demand.

Origin-destination demands are estimated 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 $s \in \mathcal{S}$ *(items/area*^{2} $*$ *time)*. Let $\lambda_i^s(x)$ denote the trip attraction rate (inbound flow) in a region of unit area about x *(items/area*$*$*time)* where $\lambda_i^s(x) = \int_{x^o \in \mathcal{A}} \lambda^s(x^o, x) dx^o$. Let $\lambda_o^s(x)$ denote the trip generation rate (outbound flow) about x *(items/area*$*$*time)* where $\lambda_o^s(x) = \int_{x^i \in \mathcal{A}} \lambda^s(x, x^i) dx^i$.

Exact customer location data are estimated with spatial customer densities for service level, $\delta^s(x)$ *(customers/area)*. The number of points in a subregion $\mathcal{A}'$ of $\mathcal{A}$ is

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

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

$$\mathcal{N}_{\mathcal{A}'} \approx \delta^s(x) |\mathcal{A}'|$$

where $|\mathcal{A}'|$ denotes the area of subregion $\mathcal{A}'$.

2.3 Decision variables

Discrete location variables are replaced with continuous variables that specify the density of terminals of type $y \in \mathcal{T}$ within a region, $\Delta_y(x)$ (*terminals/area*). Additional decision variables describe item and vehicle routings. Indices $\mathcal{B} = \{i, o\}$ define trips inbound to and outbound from a terminal,

and $\mathcal{V} = \{a, t\}$ define vehicle type (air and truck).[2] A set of variables $\{h_l^{m,b}(x)\}$ defines the route headways for distribution level $l \in \mathcal{L}$ for network $m = A, G$ in direction $b \in \mathcal{B}$. 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 between an origin (either a customer or a terminal) to the region in which the stops (either terminals or customers) are located.

2.4 Service level constraints

A series of service level parameters enforces express and deferred deadlines. The maximum headway length for a route of type $l \in \mathcal{L}$ 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 $l \in \mathcal{L}$ by introducing an upper bound N_l^m. As is typical in time-sensitive package delivery operations, the length of a delivery route is often dictated by time constraints (number of stops) and not by physical volume. A maximum airport service radius, ρ, is introduced to meet air time restrictions.

2.5 Additional assumptions

Sorting costs at facilities are approximated as constants independent of all decision variables, as shown in [31]. 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 [10]. Therefore, ground vehicles will reach either their physical capacity or their maximum number of stops. It is assumed that the long haul air network is optimally configured for express items only (i.e., airport locations are determined by express demand only which, in turn, specifies available excess capacity). It is assumed that as many deferred items as possible are shifted to air in integrated networks. This

[2] 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.

is reasonable for large ground networks, as confirmed in [31]. The fractional shift in direction $b \in \mathcal{B}$ 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.

$$r_l^m(x) \approx \frac{2}{3}(\pi \Delta_y(x))^{-\frac{1}{2}} \tag{1}$$

This estimate is on the low side, but quite accurate for desirable terminal arrangements; see [10].

3 Continuum approximation: logistic cost functions

For a specific pair of origin-destination regions, the average cost per item per unit time, z, is comprised of the following components:

$$z = z_{local} + z_{access} + z_{longhaul} + z_{reposition} + z_{CT} + z_{airport} + z_{BBT} + z_{hub} \tag{2}$$

Equation (2) contains transportation costs for each distribution level: z_{local}, z_{access}, and $z_{longhaul}$; terminal costs: z_{CT}, $z_{airport}$, z_{BBT}, and z_{hub}; and vehicle repositioning costs: $z_{reposition}$. 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[3]:

c_d^u cost of overcoming distance, for vehicle of type $u \in \mathcal{V}$ ($/*distance*)

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

c_q^u 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*time*)

[3]For ease of illustration, facilities are shown to have the same costs; costs do vary by facility type in Section 5.

In addition, the following auxiliary functions are used:

$\Lambda_b^A(x)$ 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 $b = i, o$

$\Lambda_T^m(x)$ bidirectional network-specific demand, $\Lambda_T^m(x) = \sum_{b\in\mathcal{B}} \Lambda_b^m(x)$, for $m = A, G$

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

$\Lambda_T(x)$ bidirectional demand for combined networks, $\Lambda_T(x) = \sum_{b\in\mathcal{B}} \Lambda_b(x)$

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

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.

We introduce a function $f(r, v, n, \delta)$ to designate the average unit cost for items delivered in batches of size v on vehicle routing problem (VRP) routes making n stops to customers of density δ and r distance units away from a depot. It is known (see, e.g., [10]) that:

$$f(r, v, n, \delta) \approx c_d'^u + \frac{r c_d^u + c_q^u}{nv} + \left(\frac{n-1}{n}\right)\left(\frac{c_d^u k(\delta)^{-\frac{1}{2}} + c_q^u}{v}\right) \tag{3}$$

where k is a constant dependent on the distance metric; $k \approx 0.8$ for grids[4]. The first component of (3) represents the per-item handling cost, independent of distance. 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.

[4]The parameter k can be estimated through simulation, see [10].

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

$$z_{local}^{m,b}(x) = f(r_0^m(x), v_0^{m,b}(x), n_0^{m,b}(x), \delta^m(x)) \tag{4a}$$

subject to:

$$n_0^{m,b}(x) v_0^{m,b}(x) \leq V_0^m \tag{4b}$$

$$1 \leq n_0^{m,b}(x) \leq N_0^m \tag{4c}$$

$$h_0^{m,b}(x) \leq H_0^m \tag{4d}$$

$$v_0^{m,b}(x) = \frac{\Lambda_b^m(x)}{\delta^m(x)} h_0^{m,b}(x) \tag{4e}$$

$$r_0^m(x) = \frac{2}{3}(\pi \Delta_C(x))^{-\frac{1}{2}} \tag{4f}$$

$$r_0^m(x), v_0^{m,b}(x), h_0^{m,b}(x) > 0 \tag{4g}$$

Equation (4b) ensures that loads do not exceed vehicle capacity. Equation (4c) 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.[5] Equation (4d) ensures that customers are visited with a minimum frequency. Equation (4e) expresses the shipment size $v_0^{m,b}(x)$ as a function of the demand accumulated during a headway using Little's formula ($\delta^G(x) = \delta^D(x)$ and $\delta^A(x) = \delta^E(x)$ for non-integrated networks; $\delta^m(x) = \delta(x)$ for integrated networks). Equations (4d) and (4e), combined, limit the shipment sizes. Equation (4f) expresses the dependence between linehaul distance and terminal density as stated in equation (1).

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 networks, four copies of $z_{local}^{m,b}(x)$ appear in expression (2). With integrated networks, only two copies of $z_{local}^{b}(x)$ (one for each direction) appear. The same is true of constraints.

[5] Often the number of stops limits 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:

$$z_{access}^{m,b}(x) = f(r_1^m(x), v_1^{m,b}(x), n_1^{m,b}(x), \Delta_C(x)) \tag{5a}$$

subject to:

$$n_1^{m,b}(x) v_1^{m,b}(x) \leq V_1^m \tag{5b}$$

$$1 \leq n_1^{m,b}(x) \leq N_1^m \tag{5c}$$

$$h_1^{m,b}(x) \leq H_1^m \tag{5d}$$

$$v_1^{m,b}(x) = \frac{\Lambda_b^m(x)}{\Delta_C(x)} h_1^{m,b}(x) \tag{5e}$$

$$r_1^G(x) = \frac{2}{3}(\pi \Delta_B(x))^{-\frac{1}{2}} \qquad r_1^A(x) = \frac{2}{3}(\pi \Delta_P(x))^{-\frac{1}{2}} \tag{5f}$$

$$r_1^m(x), v_1^{m,b}(x), h_1^{m,b}(x) > 0 \tag{5g}$$

For integrated networks, one must specify the network demand rates that appear in (5e) with another equation since the network demand rates no longer equal the service level demand rates. Recall that $\Lambda_b^A(x) = \lambda_b^E(x) + \omega_b(x)\lambda_b^D(x)$ and $\Lambda_b^G(x) = (1 - \omega_b(x))\lambda_b^D(x)$ for $b \in \mathcal{B}$, 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 ($h_2^A = \tilde{h} = 1$ 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 [31], 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.

$$z_{longhaul}^{A,b}(x) = f(r_2^A(x), v_2^{A,b}(x), n_2^{A,b}(x), \Delta_P(x)) \tag{6a}$$

subject to:

$$n_2^{A,b}(x) v_2^{A,b}(x) \leq V_2^A \tag{6b}$$

$$1 \leq n_2^{A,b}(x) \leq N_2^A \tag{6c}$$

$$v_2^{A,b}(x) = \frac{\Lambda_b^A(x)}{\Delta_P(x)} \tilde{h} \tag{6d}$$

$$v_2^{A,b}(x) > 0 \tag{6e}$$

$$\Delta_P(x) \geq \frac{1}{\rho^2 \pi} \tag{6f}$$

Constraint (6f) ensures that the service radius from an airport does not exceed a maximum distance ρ to guarantee the timely completion of access and local tours.

With integrated routing, a fraction of deferred items may travel by air, provided excess capacity exists. Constraint (6g) is added to restrict the amount shifted $\omega_b(x)$ by the available capacity.

$$\frac{n_2^{A,b}(x)\tilde{h}}{\Delta_P(x)} \left(\nu \omega_b(x) \lambda_b^D(x) + \lambda_b^E(x)\right) \leq V_2^A, \qquad \text{for } \nu \geq 1, \omega_b(x) \leq 1 \tag{6g}$$

The constant ν 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

Since the ground network contains multiple breakbulk terminals, 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. In [10], it is shown that the

vehicle distance traveled between breakbulk terminals can be estimated without specifying the exact routing of items when vehicles travel full. Vehicles can travel full either by increasing headways or visiting multiple terminals $\left(n_2^G(x)v_2^G(x) = V_2^G\right)$. Further, as the number of terminals in the service region $\left(\int_{x\in\mathcal{A}} \Delta_B(x)dx\right)$ 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. The detour component associated with multiple stops is estimated by:

$$\left(\frac{n_2^G(x)-1}{n_2^G(x)}\right)\left(\frac{c_d k(\Delta_B(x))^{-\frac{1}{2}}+c_q}{v_2^G(x)}\right) \tag{7a}$$

Items also incur a cost $c_d^{'t}$ for handling. Shipment size is approximated as follows. The average ground network demand rate for the service region across all origins and destinations is $\bar{\Lambda}^G = \int_{x^o\in\mathcal{A}}\int_{x^i\in\mathcal{A}} \frac{\Lambda^G(x^o,x^i)}{|\mathcal{A}|^2}dx^i dx^o$ where $\Lambda^G(x,x^i)$ is the deferred demand rate: $\lambda^D(x,x^i)$ reduced by the amount shifted to air. The average breakbulk terminal density is $\bar{\Delta}_B = \int_{x\in\mathcal{A}} \frac{\Delta_B(x)}{|\mathcal{A}|}dx$. The average shipment size over all destinations collected from a breakbulk terminal at x is approximated by

$$v_2^G(x) \approx \frac{\bar{\Lambda}^G|\mathcal{A}|h_2^G(x)}{\Delta_B(x)}. \tag{7b}$$

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, $|\Lambda_o^m(x) - \Lambda_i^m(x)|$. For vehicles with capacity V

operating in a region of terminals with density $\Delta(x)$, the number of trips is $\frac{|\Lambda_o^m(x)-\Lambda_i^m(x)|}{\Delta(x)V}$. 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, $\frac{\Lambda_T^m(x)}{\Delta(x)}$, and replacing $r(x)$ with $\frac{2}{3}(\pi\Delta(x))^{-\frac{1}{2}}$, the average cost per item is:

$$\frac{\frac{2}{3}c_d|\Lambda_o^m(x)-\Lambda_i^m(x)|}{\Lambda_T^m(x)V}\left(\pi\Delta(x)\right)^{-\frac{1}{2}}. \tag{8}$$

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 [13]. Therefore, long haul repositioning is not included in the optimization phase, but is added to the total system cost. Repositioning costs differ for integrated networks due to the demand shift from ground to air.

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, $Q(x)=\frac{\Lambda_i^m(x)}{\Delta_y(x)}+\frac{\Lambda_o^m(x)}{\Delta_y(x)}$. The cost per item for a terminal is

$$g\big(Q(x),h_o(x),h_i(x)\big)=c_f'+\frac{c_f}{Q(x)}+\sum_{b=i,o}c_h h_b(x) \tag{9}$$

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 [10].

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. We present the expression for an integrated network; thus, no network subscript is needed for local operations.

$$
\begin{aligned}
\min z = \int_{x\in\mathcal{A}} \Bigg\{ & \sum_{b\in\mathcal{B}} \Lambda_b(x) \left(c_d^{'t} + \frac{r_0(x)c_d^t + c_q^t}{n_0^b(x)v_0^b(x)} + \left(\frac{n_0^b(x)-1}{n_0^b(x)}\right)\left(\frac{c_d^t k(\delta(x))^{-\frac{1}{2}} + c_q^t}{v_0^b(x)}\right)\right) \\
& + \sum_{b\in\mathcal{B}} \sum_{m=A,G} \Lambda_b^m(x) \left(c_d^{'t} + \frac{r_1^m(x)c_d^t + c_q^t}{n_1^{m,b}(x)v_1^{m,b}(x)} + \left(\frac{n_1^{m,b}(x)-1}{n_1^{m,b}(x)}\right)\left(\frac{c_d^t k(\Delta_C(x))^{-\frac{1}{2}} + c_q^t}{v_1^{m,b}(x)}\right)\right) \\
& + \sum_{b\in\mathcal{B}} \Lambda_b^A(x) \left(c_d^{'a} + \frac{r_2^A(x)c_d^a + c_q^a}{n_2^{A,b}(x)v_2^{A,b}(x)} + \left(\frac{n_2^{A,b}(x)-1}{n_2^{A,b}(x)}\right)\left(\frac{c_d^a k(\Delta_P(x))^{-\frac{1}{2}} + c_q^a}{v_2^{A,b}(x)}\right)\right) \\
& + \Lambda_o^G(x) \left(c_d^{'t} + \left(\frac{n_2^G(x)-1}{n_2^G(x)}\right)\left(\frac{c_d^t k(\Delta_B(x))^{-\frac{1}{2}} + c_q^t}{v_2^G(x)}\right)\right) \\
& + \frac{\frac{2}{3}|\Lambda_o(x)-\Lambda_i(x)|}{V_0^G\sqrt{\pi\Delta_C(x)}}c_d^t + \frac{\frac{2}{3}|\Lambda_o^A(x)-\Lambda_i^A(x)|}{V_1^G\sqrt{\pi\Delta_P(x)}}c_d^t + \frac{\frac{2}{3}|\Lambda_o^G(x)-\Lambda_i^G(x)|}{V_1^G\sqrt{\pi\Delta_B(x)}}c_d^t \\
& + \Lambda_T(x)c_f' + \Delta_C(x)c_f + \sum_{b\in\mathcal{B}} c_h\Lambda_b(x)h_0^b(x) + \Lambda_T^G(x)c_f' + \Delta_B(x)c_f \\
& + \Lambda_T^A(x)c_f' + \Delta_P(x)c_f + \sum_{b\in\mathcal{B}} c_h\Lambda_b^G(x)h_1^{G,b}(x) + \Lambda_o^G(x)c_h h_2^G(x) + \sum_{b\in\mathcal{B}} c_h\Lambda_b^A(x)\left(\tilde{h} + h_1^{A,b}(x)\right)\Bigg\} dx
\end{aligned}
\tag{10}
$$

The integrand of (10) 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 (10) 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 [31]. 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 $\gamma_b^s(x), s \in \mathcal{S}, b \in \mathcal{B}$ $(items)$. 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:

$$\theta^{A,b} V_2^A + 3\sqrt{\theta^{A,b} \gamma_b^E V_2^A} \leq V_2^A \tag{11}$$

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 (6g) 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.[6] The vehicle repositioning cost due to stochastic effects alone can be approximated as a transportation problem, as shown in [13], and added to deterministic repositioning costs.[7] The average number of total empty vehicle miles across many days required to reposition vehicles at the least cost each day is a function of the total area of the service region $\mathcal{A}$, the number of terminals, $\bar{\Delta}_y(x)|\mathcal{A}|$, and $\sigma_y(x)$, where $\sigma_y(x)$ is the standard deviation of the flows inbound to and outbound from a terminal equal to $\sqrt{\frac{\gamma_i^G(x)\Lambda_i^G(x) + \gamma_o^G(x)\Lambda_o^G(x)}{\Delta_y(x)V}}$. The stochastic repositioning cost per terminal per unit time is:

$$z_{reposition} = c_d^t \sigma_y(x) \Delta_y(x)^{-\frac{1}{2}} \left(1 + 0.078 \log_2(\bar{\Delta}_y(x)|\mathcal{A}|)\right) \tag{12}$$

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

[6] 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.

[7] 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.

4 Solution method

This section describes how the problem can be separated into a series of subproblems that can be solved easily. First (10) 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 (4b) or (4c) must be binding (on the upper side). The same is true for (5b) and (5c), and for (6b) and (6c). Therefore, we can replace $n_l^{m,b}(x)$ with $\min\left\{N_l^m, \frac{V_l^m}{v_l^{m,b}(x)}\right\}$. Equations (4d), (5d), and (6d) are used to eliminate $v_l^{m,b}(x)$ and equations (4e) and (5e) to eliminate $r_l^m(x)$. Recall that $\omega_b(x)$ can be set equal to the largest possible value consistent with (6g) 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 α, β, χ, ψ, and Π 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:

$$\begin{aligned}\min z = \int_{x\in\mathcal{A}} \Bigg\{ & \sum_{b=i,o} \left(\alpha_1^b h_0^b(x) + \alpha_2 (h_0^b(x))^{-1}\right) + \beta_1 \Delta_C(x)^{-\frac{1}{2}} \\ & + \sum_{b=i,o}\sum_{m=A,G} \left(\beta_2 \frac{\sqrt{\Delta_C(x)}}{h_1^{m,b}(x)} + \beta_3 \frac{\Delta_C(x)}{h_1^{m,b}(x)} + \beta_4^b h_1^{m,b}(x)\right) + \beta_6 \Delta_C(x) \\ & + \chi_1 \Delta_P^{-\frac{1}{2}}(x) + \chi_2 \Delta_P(x) + \chi_3 \Delta_P^{\frac{1}{2}}(x) \\ & + \psi_1 \Delta_B^{-\frac{1}{2}}(x) + \psi_2 \Delta_B(x) + \psi_3 \frac{\Delta_B(x)^{\frac{1}{2}}}{h_2^G(x)} + \psi_4 \frac{\Delta_B(x)}{h_2^G(x)} + \psi_5 h_2^G(x) + \Pi \Bigg\} dx \end{aligned} \quad (13a)$$

subject to:

$$\frac{\Lambda_b(x)\delta(x)}{N_0V_0} \leq h_0^b(x) \leq \frac{\Lambda_b(x)\delta(x)}{V^0} \qquad \forall b \in \mathcal{B} \tag{13b}$$

$$\frac{\Lambda_b^m(x)}{V_1^m} \leq \frac{\Delta_C(x)}{h_1^{m,b}(x)} \leq \frac{N_1^m\Lambda_b^m(x)}{V_1^m} \qquad \forall b \in \mathcal{B}; m = A, G \tag{13c}$$

$$\frac{\Lambda_b^A(x)}{V_2^A} \leq \Delta_P(x) \leq \frac{N_2^A\Lambda_b^A(x)}{V_2^A} \qquad \forall b \in \mathcal{B} \tag{13d}$$

$$\frac{\bar{\Lambda}^G|\mathcal{A}|}{V_2^G} \leq \frac{\Delta_B(x)}{h_2^G(x)} \leq \frac{N_2^G\bar{\Lambda}^G|\mathcal{A}|}{V_2^G} \tag{13e}$$

$$0 < h_l^{m,b}(x) \leq H_l^m \qquad \forall b \in \mathcal{B}; m = A, G; l \in \mathcal{L} \tag{13f}$$

$$\Delta_P(x) \geq \frac{1}{\rho^2\pi} \tag{13g}$$

Note that (13) can be decomposed into five classes of subproblems that involve the following groups of decision functions:

1^o	local outbound headways, $h_0^o(x)$
1^i	local inbound headways, $h_0^i(x)$
2	consolidation terminal densities and access headways, $\Delta_C(x), h_1^{m,b}(x)$
3	airport densities, $\Delta_P(x)$
4	breakbulk terminal densities and long haul ground headways, $\Delta_B(x), h_2^G(x)$.

The subproblems in each class are analyzed below.

4.1 Subproblems 1^o and 1^i: local headways

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

$$\min z_{1^b} = \int_{x\in\mathcal{A}} \left(\alpha_1^b h_0^b(x) + \frac{\alpha_2}{h_0^b(x)}\right) dx \tag{14a}$$

subject to:

$$\frac{\Lambda_b(x)\delta(x)}{N_0V_0} \leq h_0^b(x) \leq \frac{\Lambda_b(x)\delta(x)}{V_0} \tag{14b}$$

$$0 < h_0^b(x) \leq H_0 \tag{14c}$$

Note that (14) can be further decomposed by x because the integrand and the constraints are local in nature. Hence, one can simply minimize the integrand of (14a) for every x and sum across all 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: $h_0^b(x)^* = \sqrt{\frac{\alpha_2}{\alpha_1^b}}$ if this value satisfies the constraints. Otherwise, it is one of the extreme points defined by the constraints. The result should be intuitive, since with higher transportation costs, headways are lengthened, and with higher rent costs, headways are shortened. We find that for reasonable values of the parameters, (14c) is binding at its upper bound.

4.2 Subproblem 2: consolidation terminal densities and access headways

Subproblem 2 is:

$$\min z_2 = \int_{x\in\mathcal{A}} \Bigg(\beta_1 \Delta_C(x)^{-\frac{1}{2}} + \sum_{b=i,o} \sum_{m=A,G} \left(\beta_2 \frac{\sqrt{\Delta_C(x)}}{h_1^{m,b}(x)} + \beta_3 \frac{\Delta_C(x)}{h_1^{m,b}(x)} + \beta_4^b h_1^{m,b}(x) \right) + \beta_6 \Delta_C(x) \Bigg) dx \quad (15a)$$

subject to:

$$\frac{\Lambda_b^m(x)}{V_1^m} \leq \frac{\Delta_C(x)}{h_1^{m,b}(x)} \leq \frac{N_1^m \Lambda_b^m(x)}{V_1^m} \qquad b \in \mathcal{B}; m = A, G \quad (15b)$$

$$0 < h_1^{m,b}(x) \leq H_1^m \qquad b \in \mathcal{B}; m = A, G \quad (15c)$$

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 (15) into a convex problem with linear constraints: $w_C(x) = \ln(\Delta_C(x))$, $w_{m,b}(x) = \ln(h_1^{m,b}(x)), b \in \mathcal{B}; m = A, G$. The transformed problem is:

$$\min z_{2'} = \int_{x\in\mathcal{A}} \Bigg(\beta_1 e^{-\frac{w_C(x)}{2}} + \sum_{b=i,o} \sum_{m=A,G} \left(\beta_2 e^{\frac{w_C(x)}{2} - w_{m,b}(x)} + \beta_3 e^{w_C(x) - w_{m,b}(x)} + \beta_4^b e^{w_{m,b}(x)} \right) + \beta_6 e^{w_C(x)} \Bigg) dx \quad (16a)$$

subject to:

$$\ln\left(\frac{\Lambda_b^m(x)}{V_1^m}\right) \leq w_C(x) - w_{m,b}(x) \leq \ln\left(\frac{N_1^m \Lambda_b^m(x)}{V_1^m}\right) \qquad b \in \mathcal{B}; m = A, G \tag{16b}$$

$$w_{m,b}(x) \leq \ln(H_1^m) \qquad b \in \mathcal{B}; m = A, G \tag{16c}$$

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:

$$\min z_{3'} = \int_{x \in \mathcal{A}} \left(\chi_1 e^{-\frac{w_P(x)}{2}} + \chi_2 e^{w_P(x)} + \chi_3 e^{\frac{w_P(x)}{2}} \right) dx \tag{17a}$$

subject to:

$$\ln\left(\frac{\Lambda_b^A(x)}{V_2^A}\right) \leq w_P(x) \leq \ln\left(\frac{N_2^A \Lambda_b^A(x)}{V_2^A}\right) \qquad b \in \mathcal{B} \tag{17b}$$

$$w_2(x) \geq \ln(\frac{1}{\rho^2 \pi}) \tag{17c}$$

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:

$$\min z_{4'} = \int_{x \in \mathcal{A}} \left(\psi_1 e^{-\frac{w_B(x)}{2}} + \psi_2 e^{w_B} + \psi_3 e^{\frac{w_B(x)}{2} - w_2(x)} + \psi_4 e^{w_B(x) - w_2(x)} + \psi_5 e^{w_2} \right) dx \tag{18a}$$

subject to:

$$\ln\left(\frac{\bar{\Lambda}^G|\mathcal{A}|}{V_2^G}\right) \leq w_B(x) - w_2(x) \leq \ln\left(\frac{N_2\bar{\Lambda}^G|\mathcal{A}|}{V_2^G}\right) \tag{18b}$$

$$w_2(x) \leq \ln(H_2^G) \tag{18c}$$

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.

5 Case study

The design methodology introduced in Sections 3 and 4 is used to obtain network configurations for large package delivery networks roughly the size of the contiguous United States. Operating costs and statistics are derived from [23] and company literature. Population density is used as a proxy for package demand rates (λ) and housing density as a proxy for customer density (δ). 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 [34]. 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 mi^2. The areas of the twenty subregions range from 16,383 mi^2 to 225,974 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 3, 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.

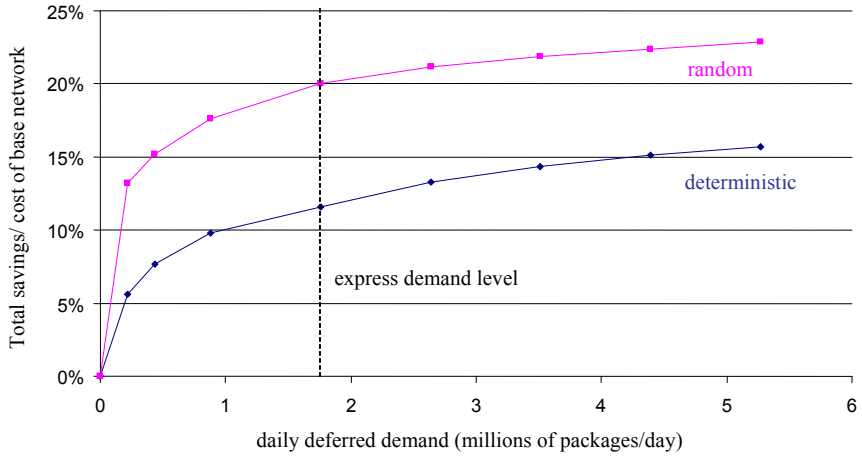

Figure 3: Cost savings from integration

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. 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 due to the geographic distribution of existing facilities, 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 re-designing 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.

5.1 Model accuracy

As discussed in Section 2, we propose the use of continuum approximations for network design problems in cases where discrete problems can not be fully formulated due to operating complexity or problem size. Therefore, in such cases the accuracy of the complete design methodology cannot be validated with a direct comparison of discrete and continuum formulations. As a result, component cost models are validated individually here, and complemented by validation tests from the literature. Since the input data to planning models are often continuous, tests in the literature compare the continuum average cost predictions with the average cost produced by discrete optimization models across many samples of data simulated from the continuous data. Vehicle routing cost formulae have been validated in [16] and [30]. 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 [5] and [28]. Long haul operating costs are validated using a discrete formulation from [33] given a fixed network of terminals. The difficulties encountered when solving large problem instances highlight the complexity of the discrete formulation. Empty vehicle repositioning costs are validated with TSP simulations in [13]. Terminal cost approximations are validated in [31]. 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 from our model and the optimum of a discrete version of the same problem cannot exceed 5% and is likely much smaller.

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. 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 continuum 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 needed 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 then be used to geographically separate discrete terminal location problems and vehicle and item routing problems, as shown in [31]. Algorithms to translate continuum approximation outputs into discrete solutions are developed in [28].

A key value of the continuum approximation method is its flexibility. While the problem presented in this paper is quite stylized, many extensions can be envisioned that are beyond the scope of this paper. For example, [31] considers hybrid levels of network integration in which facilities are shared among service levels and modes, but routing remains separate. Additionally, one could allow a fraction of express items to fly directly between airports when demand and distance dictate, see [21]. Issues related to design of the air network are explored in [31].

Acknowledgments

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

References

[1] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms and Applications.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1993.

[2] Andrew Armacost. *Composite Variable Formulations for Express Shipment Service Network Design.* PhD dissertation, Massachusetts Institute of Technology, September 2000.

[3] M.O. Ball, T.L. Magnanti, C.L. Monma, and G.L. Nemhauser, editors. *Network Models*, volume 7 of *Handbooks in Operations Research and Management Science*. Elsevier Science Publishing, New York, 1995.

[4] Cynthia Barnhart and Rina R. Schneur. Air network design for express shipment service. *Operations Research*, 44(6):852–863, 1996.

[5] James F. Campbell. Continuous and discrete demand hub location problems. *Transportation Research B*, 27B(6):473–482, 1993.

[6] G. Clarens and V.F. Hurdle. An operating strategy for a commuter bus sytems. *Transportation Science*, 9:1–20, 1975.

[7] Teodor Crainic. Service network design in freight transportation. *European Journal of Operational Research*, 122(2):272–288, 2000.

[8] Carlos F. Daganzo. Modeling distribution problems with time windows: Part I. *Transportation Science*, 21(3):171–179, 1987.

[9] Carlos F. Daganzo. Modeling distribution problems with time windows: Part II. *Transportation Science*, 21(3):180–187, 1987.

[10] Carlos F. Daganzo. *Logistics Systems Analysis*. Springer, New York, 1999.

[11] Carlos F. Daganzo and Alan L. Erera. On planning and design of logistics systems for uncertain environments. In M. Grazia Speranza and Paul Stahly, editors, *New Trends in Distribution Logistics*, pages 100–105. Springer, Berlin, 1999.

[12] Carlos F. Daganzo and Gordon F. Newell. Configuration of physical distribution networks. *Networks*, 16:113–132, 1986.

[13] Carlos F. Daganzo and Karen R. Smilowitz. Bounds and approximations for the transportation problem of linear programming and other scalable networks. *Transportation Science*, 38(3):343–356, 2004.

[14] Mark S. Daskin. Logistics: An overview of the state of the art and perspectives on future research. *Transportation Research*, 19A(5/6):383–398, 1985.

[15] Samuel Eilon, C.D. T. Watson-Gandy, and Nicos Christofides. *Distribution Management*. Griffin, London, 1971.

[16] Alan L. Erera. *Design of Logistics Systems for Uncertain Environments*. PhD dissertation, University of California, Berkeley, Institute of Transportation Studies, 2000.

[17] A.M. Geoffrion. The purpose of mathematical programming is insight, not numbers. *Interfaces*, 7(1):81–92, 1976.

[18] Randolph Hall. Dispatching regular and express shipments between a supplier and manufacturer. *Transportation Research B*, 23B(3):195–211, 1989.

[19] Randolph Hall. Characteristics of multi-stop / multi-terminal delivery routes with backhauls and unique items. *Transportation Research B*, 25B(6):391–403, 1991.

[20] Anthony Fu-Wha Han. *One-to-Many Distribution of Nonstorable Items: Approximate Analytical Models*. PhD dissertation, University of California, Berkeley, 1984.

[21] C.Y. Jeng. *Routing strategies for an idealized airline network*. PhD dissertation, University of California, Berkeley, 1987.

[22] W.C. Jordan and L.D. Burns. Truck backhauling on two terminal networks. *Transportation Research B*, 18B(6):487–503, 1984.

[23] Max Karl Kiesling. *A comparison of freight distribution costs for combination and dedicated carriers in the air express industry*. PhD dissertation, University of California, Berkeley, 1995.

[24] Andre Langevin, Pontien Mbaraga, and James F. Campbell. Continuous approximation models in freight distribution: An overview. *Transportation Research B*, 30B(3):163–188, 1996.

[25] T.L. Magnanti and R.T. Wong. Network design and transportation planning: Models and algorithms. *Transportation Science*, 18(1):1–55, 1984.

[26] G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization*. Wiley, New York, 1999.

[27] G.F. Newell. Scheduling, location, transportation, and continuum mechanics: some simple approximations to optimization problems. *SIAM, Journal of Applied Mathematics*, 25:346–360, 1973.

[28] Yanfeng Ouyang and Carlos F. Daganzo. Discretization and validation of the continuum approximation scheme for terminal system design. Technical Report UCB-ITS-WP-2003-2, Institute of Transportation Studies, University of California at Berkeley, 2003.

[29] W.B. Powell and Y. Sheffi. The load planning problem of motor carriers: Problem description and a proposed solution approach. *Transportation Research A*, 17A(6):471–480, 1983.

[30] Francesc Robuste, Carlos F. Daganzo, and Reginald R. Souleyrette II. Implementing vehicle routing models. *Transportation Research B*, 24(4):263–286, 1990.

[31] Karen Smilowitz. *Design and Operation of Multimode, Multiservice Logistics Systems.* PhD dissertation, University of California, Berkeley, Institute of Transportation Studies, 2001.

[33] Karen Smilowitz, A. Atamtürk, and Carlos F. Daganzo. Deferred item and vehicle routing within integrated networks. *Transportation Research. Part E, Logistics and Transportation Review*, 39:305–323, 2003.

[32] Karen Smilowitz and Carlos F. Daganzo. Cost modeling and solution techniques for complex transportation systems. IEMS Working Paper 04-006, 2003.

[34] U.S. Census Bureau. Tiger/geographic identification code scheme, 1990.

Appendix: Notation

Network sets

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

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

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

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

$\mathcal{T}$ Set of terminal (node) types, $\mathcal{T} = \{C, B, P, H\}$ 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/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*time)*

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

$\lambda_o^s(x)$ trip generation rate about x *(items/area*time)*; $\lambda_o^s(x) = \int_{x \in \mathcal{A}} \lambda^s(x^o, x) dx$

Level of Service Parameters

H_l^m maximum headway length for network $m = A, G$ for a route of type $l \in \mathcal{L}$

$\tilde{h}$ restricted operating headway in long haul air network, $\tilde{h} = 1$ day

N_l^m maximum number of stops for network $m = A, G$ on a route of type $l \in \mathcal{L}$

V_l^m vehicle capacity for network $m = A, G$ for a route of type $l \in \mathcal{L}$ ($items$)

ρ maximum airport service radius *(distance)*

Cost Parameters

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

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

c_q^u 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*time$)

Decision functions

$\Delta_y(x)$ density of terminals of type $y \in \mathcal{T}$ *(terminals/area)*

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

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

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

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

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

Auxiliary demand functions

$\Lambda_b^A(x)$ 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 $b = i, o$

$\Lambda_T^m(x)$ bidirectional network-specific demand, $\Lambda_T^m(x) = \sum_{b \in \mathcal{B}} \Lambda_b^m(x)$, for $m = A, G$

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

$\Lambda_T(x)$ bidirectional demand for combined networks, $\Lambda_T(x) = \sum_{b \in \mathcal{B}} \Lambda_b(x)$

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

Coefficients and constants

Constant Π, where it is understood that Π is a function of x.

$$\begin{aligned}\Pi =& \Lambda_T(x)\left(c_d^{t'} - \frac{c_d^t k(\delta(x))^{-\frac{1}{2}}}{V_0}\right) + \Lambda_T(x)c_d^{t'} + \Lambda_o^G(x)c_d^{t'} + \Lambda_T^A(x)\left(c_d^{a'} + 2c_h\tilde{h}\right) \\ &+ \Lambda_o(x)c_k\log(2) + 2\Lambda_T(x)c_f'\end{aligned}$$

Coefficients for local operating headways:

$$\alpha_1^b = \Lambda_b c_h;\ b = i, o \qquad\qquad \alpha_2 = c_d^t k(\delta)^{\frac{1}{2}} + c_q^t\delta$$

Coefficients for consolidation terminal densities and access operating headways:

$$\beta_1 = \Lambda_T\left(\frac{\frac{2}{3}c_d^t}{V_0}\pi^{-\frac{1}{2}} - \frac{c_d^t k}{V_1}\right) + \frac{\frac{2}{3}|\Lambda_o - \Lambda_i|}{V_0}\pi^{-\frac{1}{2}}c_d^t$$

$$\beta_2 = c_d^t k \qquad \beta_3 = c_q^t \qquad \beta_4^b = \Lambda_b^A\frac{c_h}{2} \qquad \beta_5^b = \Lambda_b^G c_h;\ b = i, o \qquad \beta_6 = c_f$$

Coefficients for airport densities:

$$\chi_1 = \Lambda_T^A \left(\frac{\frac{2}{3} c_d^t}{V_1^A} \pi^{-\frac{1}{2}} \right) + \frac{\frac{2}{3} |\Lambda_o^A - \Lambda_i^A|}{V_1^A} \pi^{-\frac{1}{2}} c_d^t \quad \chi_2 = c_f + \sum_{b \in \mathcal{B}} \frac{r_2^A c_d^t + n_2^{A,b} c_q^t}{n_2^{A,b}} \quad \chi_3 = \sum_{b \in \mathcal{B}} \frac{n_2^{A,b} - 1}{n_2^{A,b}} c_d^t k$$

Coefficients for breakbulk terminal densities and long haul operating headways:

$$\psi_1 = \Lambda_o^G \left(\frac{\frac{2}{3} c_d^t}{V_1^G} \pi^{-\frac{1}{2}} \right) + \frac{\frac{2}{3} |\Lambda_o^G - \Lambda_i^G|}{V_1^G} \pi^{-\frac{1}{2}} c_d^t - \bar{\Lambda}^G c_d^t \frac{k}{V_2^G}$$

$$\psi_2 = c_f \qquad \psi_3 = c_d^t k \qquad \psi_4 = c_q^t \qquad \psi_5 = \Lambda_o^G c_h$$