Flexible Milk-Runs for Stochastic Vehicle Routing

Ben Peterson  
*Carnegie Mellon University*

Willem-Jan van Hoeve  
*Carnegie Mellon University, vanhoeve@andrew.cmu.edu*

Sham Kekre  
*Carnegie Mellon University, kekresh@cmu.edu*

Laurens Debo  
*University of Chicago*

Follow this and additional works at: [http://repository.cmu.edu/tepper](http://repository.cmu.edu/tepper)
Flexible Milk-Runs for Stochastic Vehicle Routing

Ben Peterson, Willem-Jan van Hoeve, Sham Kekre
Tepper School of Business, Carnegie Mellon University, Pittsburgh, USA
benp@cmu.edu, vanhoeve@andrew.cmu.edu, kekresh@cmu.edu

Laurens Debo
University of Chicago Graduate School of Business, Chicago, USA
laurens.debo@ChicagoBooth.edu

We study a vehicle routing problem with stochastic demands in which the goal is to find an optimal set of vehicle routes, such that the capacity of each vehicle is not exceeded with a given probability. We introduce ‘flexible milk-runs’, or flex-runs, to model this problem as a set covering problem to find (near-)optimal solutions. We apply our methodology to design new freight routes for the North-American division of the Bosch/Siemens Home Appliances Corporation. Our computational experiments indicate an expected transportation cost reduction of up to 25%, while at the same time the new routes realize overall increase in robustness with respect to demand fluctuations.

Key words: stochastic vehicle routing; column generation; real-world application

History:

1 Introduction

The vehicle routing problem (VRP) is a central problem in operations research, asking to optimally route a fleet of trucks to serve a set of customers [19]. In this work, we consider a stochastic vehicle routing problem in which the customer demands are uncertain and splittable. In the variant that we consider, the goal is to design a fixed set of truck routes, such that the capacity of each truck is not exceeded with a given probability. This variant occurs often in practical industrial applications. For example, many manufacturing facilities use third-party logistics providers to transport the incoming freight from their suppliers. Even though a truck can be ordered on a per-event basis, it is common to order trucks in advance, for example by using a fixed schedule for each supplier pick-up. One benefit of applying a fixed schedule is having a stable and predictable distribution pattern. A second benefit is that a fixed schedule often allows to negotiate better rates for the manufacturing facility. One important drawback, however, is that when a truck route involves multiple suppliers,
the actual load of a truck may fluctuate more, leading to under-usage or exceeded capacity. We show in this paper that consolidating shipments can significantly reduce transportation costs, while the resulting solution can be more robust with respect to volume changes.

Our proposed solution involves the careful selection of suppliers into consolidated routes that we call flex-runs. Flex-runs are essentially flexible milk-runs (the common term for consolidated routes in logistics) designed to handle volume swings along the route. This approach corresponds to the well-known ‘set covering’ formulation that has been applied successfully to model and solve deterministic VRPs. An important aspect of shipment consolidation is to maintain flexibility with respect to variation in shipment volumes, in the case of stochastic VRPs. To account for this, we design our flex-runs in such a way that the expected volume change for one supplier offsets the expected volume change for the other suppliers. In other words, by exploiting negative correlations between the volume changes of suppliers, we are able to guarantee a lower bound (typically 90-99%) on the probability that the vehicles serving each flex-run will remain within capacity.

Despite the success of the set covering formulation, typically coupled with a delayed column generation solving process, in the context of deterministic VRPs, it is not commonly applied to stochastic VRPs. Novoa et al. [14] are the first to propose a set covering formulation for stochastic VRPs. In this work we apply the same idea, but we propose and evaluate a slightly more general model that allows splittable demands. Furthermore, we emphasize the correlation between the customer demands and explicitly exploit this information to design robust flex-runs. In addition, we propose a different recourse penalty function than was done in [14]. Finally, we present an extensive real-world case study for Bosch/Siemens Home Appliances which indicates that cost savings of up to 25% are possible using more robust routes, when compared to the original situation.

The remainder of the paper is structured as follows. We first provide a literature review in Section 2. In Section 3 we provide a detailed problem description. We present our proposed solution method using flex-runs in Section 4. Section 5 is dedicated to the case study, in which we outline the characteristics of the application, describe the data set, present computational results, and discuss the practical implications of our proposed solution. Finally, we present the main conclusions in Section 6.
2 Related Literature

The vehicle routing problem (VRP) has been extensively studied in operations research for more than 50 years [11]. Given the importance of the problem, many solution techniques have been proposed, ranging from heuristic search techniques to mixed integer linear programming and constraint programming. A detailed exposition of all these developments is beyond the purpose of this paper, and we refer to Toth and Vigo [19] for a survey of the most important results in the literature. In this section we restrict ourselves to those references that are most related to our work.

One particularly popular approach for modeling and solving deterministic VRPs is the ‘set covering’ formulation, originally proposed by Balinski and Quandt [2]. In this integer linear programming formulation, truck routes are represented by a column in the linear model, indexed by the locations that need to be visited. The total distance of the route is represented as the objective coefficient for that column. Then, an integer variable corresponding to each column is introduced to determine the number of times the column (route) is used in the solution. Solving the VRP then amounts to selecting a subset of the columns such that all locations are covered. The original paper of Balinski and Quandt [2] proposes to generate all (exponentially many) columns to a model. Nowadays, advanced delayed column generation procedures are applied instead to solve large-scale routing problems, see, e.g., Desrochers et al. [5], Barnhart et al. [3], Pajunas et al. [15]. One of the main modeling advantages of the set covering formulation is that any type of side constraint can be imposed on the routes, as long as the resulting columns are tractable to compute. We take advantage of this in our work, as we apply a set covering formulation for stochastic VRPs, where we impose probabilistic constraints on the routes. Our solution method does not apply delayed column generation, however.

Even though most studies on vehicle routing studies consider deterministic VRPs, there is a substantial stream of literature on stochastic VRPs; a survey of solution methods for stochastic VRPs is presented by Gendreau, Laporte, and Séguin [8]. Similar to general stochastic programming, there are two main stages in solving stochastic VRPs. First, an initial solution is computed based on the stochastic information. Most often, a ‘chance-constrained model’ is used for this. Second, the stochastic elements of problem are realized, and a recourse (or corrective) action is applied to the first stage solution. When the additional costs of recourse are taken into account, the problem is usually modeled as a ‘stochastic prob-
Almost all solution methods for stochastic VRPs can be characterized in this way, including the method we propose in this work.

The classical recourse strategy for stochastic VRPs is to have each vehicle follow its first-stage route, and have it make intermediate stops at the depot to unload when it reaches its capacity along the route, see Dror, Laporte, and Trudeau [6]. Variants of this strategy exist, for example by limiting the number of failures along the route [18, 16], or having the vehicle return to the depot pro-actively (before an actual failure occurs) [4]. More intricate recourse strategies involve the cooperation between vehicles, by allocating excess demand of one vehicle to another (designated) vehicle [14, 1].

Most solution methods for the stochastic VRP are inspired by solution methods for the deterministic VRP. For example, several papers adapt (meta-)heuristics for the deterministic VRP to the stochastic case, e.g., [17, 16, 18]. Another approach, taken by [17], is to convert the (chance-constrained) stochastic VRP into an equivalent deterministic VRP to be solved by any existing method. Finally, exact solution approaches are taken by [12, 7, 9, 13] using the so-called ‘integer L-shaped method’.

The paper by Novoa, Berger, Linderoth, and Storer [14] is most closely related to our work. It describes a set covering formulation, using column generation, very similar to ours, to model the stochastic VRP with uncertain demands. One of the main differences with our work is their recourse strategy: they allow cooperation between vehicles such that ‘excess customers’ of one vehicle can be served by another vehicle. That recourse action is clearly applicable to certain practical situations, however not in the case study that we consider. Instead, our recourse action is based on the observation that excess demand is often hauled by ad-hoc trucks that are dedicated to serve single location. Similar to [14], we apply a limited column generation procedure, allowing up to three locations per vehicle route. An important difference that distinguishes our work from [14] is that we provide computational results on real-world data, while explicitly exploiting correlations between customers, whereas Novoa et al. [14] report results on benchmarks from the literature in which customer demands are assumed independent. Therefore, from a computational perspective the two papers complement each other.
3 Problem Definition

We first introduce the stochastic vehicle routing problem with uncertain splittable demands for a single time period. We are given a set of ‘locations’ \( V \), in which one designated location \( v_0 \in V \) corresponds to the ‘depot’, while the other locations \( V \setminus \{v_0\} \) correspond to ‘customers’. At the depot a set of identical vehicles \( K \) is stationed, each with a given capacity \( \text{Cap} \). Each customer has an associated uncertain pickup demand that needs to be transported to the depot using one or more vehicles. The ‘distance’ (or cost) matrix is given by \( C = (c_{ij}) \) where entry \( c_{ij} \) represents the ‘distance’ between two locations \( i, j \in V \).

The demand to be picked up at each customer \( v \in V \setminus \{v_0\} \) is represented by a random variable \( \xi_v \) with mean \( \mu_v \) and variance \( \sigma_v^2 \). We assume that some probability distribution underlying \( \xi_v \) has been given, for example a normal distribution or a Gamma distribution. The covariance of the demand distributions is denoted by the matrix \( \Sigma = (\sigma_{ij}) \), where \( \sigma_{ij} \) represents the covariance between two customers \( i, j \in V \setminus \{v_0\} \).

In the most general form, we assume that the demand for each customer can be split in an arbitrary (continuous) way. For each customer \( i \in V \setminus \{v_0\} \) and truck \( k \in K \), we introduce a continuous variable \( y_{ik} \in [0, 1] \) representing the proportion of the demand of customer \( i \) that will be picked up by vehicle \( k \). We note that in most practical cases, the demand splitting is done according to a given granularity \( G \), i.e., \( G \) is a natural number representing the smallest amount of demand that can be picked up. When the demands are deterministic, one can transform the VRP with splittable demands into an equivalent VRP with unsplittable demands by creating \( W/G \) copies of each customer (one for each demand partition), where the customer demand \( W \) is a multiple of \( G \). For stochastic demands, however, this is not possible because the number of necessary copies depends on the realization of the stochastic demand. We will discuss in the second part of the paper how we handle this ‘discrete’ splitting of the demand in a practical application.

In order to mathematically represent the routes of the vehicles, we introduce a binary variable \( x_{ijk} \) representing whether vehicle \( k \) travels from location \( i \) to \( j \) \( (x_{ijk} = 1) \) or not \( (x_{ijk} = 0) \), for \( i, j \in V \) and \( k \in K \). The single-period stochastic vehicle routing problem with uncertain demands can now be stated as the following ‘chance constrained’ program. Given a probability \( \alpha \), find an allocation of customer demand to vehicles such that the expected total load of each vehicle does not exceed its capacity with probability \( \alpha \), while the total distance traveled is minimized:
\[
\min \sum_{i \in V} \sum_{j \in V} c_{ij} \sum_{k \in K} x_{ijk}
\]

\(\text{s.t.}\)
\[
\sum_{k \in K} y_{ik} = 1 \quad \forall i \in V \setminus \{v_0\} \tag{2}
\]
\[
y_{ik} \leq \sum_{j \in V} x_{jik} \quad \forall i \in V \setminus \{v_0\}, \forall k \in K \tag{3}
\]
\[
\Pr \left( \sum_{i \in V \setminus \{v_0\}} \xi_i y_{ik} \leq \text{Cap} \right) \geq \alpha \quad \forall k \in K \tag{4}
\]
\[
\sum_{k \in K} \sum_{j \in V \setminus \{v_0\}} x_{v_0 jk} = |K| \tag{5}
\]
\[
\sum_{j \in V} x_{ijk} = \sum_{j \in V} x_{jik} \quad \forall i \in V, j \in K \tag{6}
\]
\[
\sum_{i \in S} \sum_{j \in S} x_{ijk} \leq |S| - 1 \quad \forall S \subseteq V \setminus \{v_0\}, |S| \geq 2, \forall k \in K \tag{7}
\]

In this model, the objective (1) minimizes the total distance traveled by all vehicles. Constraint set (2) ensures that all demand is picked up for each customer, while constraints (3) make sure that only those trucks visiting customer \(i\) can pick up demand from \(i\), and vice-versa that a vehicle must visit \(i\) in order to pick up its demand. Constraint set (4) states that the probability that the capacity of a vehicle will not be exceeded is at least \(\alpha\). Constraints (5) make sure that each vehicle in \(K\) starts its route from the depot. Finally, the ‘flow-balance’ constraints (6) and ‘subtour elimination’ constraints (7) ensure that each vehicle follows a closed tour.

### 4 Set Covering Using Flexible Milk-Runs

We next outline our proposed solution approach for the stochastic VRP described in Section 3. It combines the well-known set-covering formulation of the VRP, using column generation, with a stochastic model for the vehicle load distributions. The first part corresponds to a chance-constrained approach, in which we find a solution that does not exceed the capacity of each truck with at least a given probability. The second part introduces a specific recourse model in which we measure the effect of repairing the solution in case we do exceed the capacity.
4.1 Flex-Run Design

The basic idea is to group locations together onto a vehicle in such a way that the joint demand distribution will not exceed the vehicle’s capacity. The routes (historically called ‘milk-runs’) that are thus formed should be flexible with respect to demand fluctuations, and we therefore refer to them as ‘flexible milk-runs’ or flex-runs. Formally, we represent a flex-run $f$ by a vector of pick-up proportions $p_f : V \setminus \{v_0\} \rightarrow [0, 1]$ for the locations. If $p_{fv} = 0$ for some location $v \in V \setminus \{v_0\}$, the vehicle corresponding to flex-run $f$ will not visit $v$. We denote the cardinality of the set of locations that $f$ visits by $|f|$, i.e., $|f| = |\{v \mid p_{fv} > 0\}|$.

With each flex-run $f$, we associate a (minimum) travel distance (or cost) $c_f$, and we assume that $f$ obeys the capacity constraint on its associated vehicle:

$$
\Pr \left( \sum_{v \in V \setminus \{v_0\}} \xi_v p_{fv} \leq \text{Cap} \right) \geq \alpha. 
$$

(8)

We regard the total demand picked up by a flex-run $f$ as a random variable $X_f$ with mean $\mu_f$ and variance $\sigma_f^2$ given by

$$
\mu_f = \sum_{v \in V \setminus \{v_0\}} p_{fv} \mu_v, \quad \sigma_f^2 = \sum_{v \in V \setminus \{v_0\}} \sum_{w \in V \setminus \{v_0\}} p_{fw}^T \Sigma p_{fw}
$$

In other words, $X_f$ accumulates the probability distributions of the demand variables $\xi_v$ for the locations $v \in V \setminus \{v_0\}$, proportional to the pick-up amounts $p_{fv}$. Constraint (8) then amounts to

$$
\Pr (X_f \leq \text{Cap}) \geq \alpha.
$$

(9)

An appropriate probability distribution for $X_f$ depends on the (historical) problem data. It may theoretically look appealing to apply a (truncated) normal distribution to $X_f$, inspired by the central limit theorem. However, as we explicitly assume that the random variables $\xi_v$ are not independent, and they moreover may assume different probability distributions, the central limit theorem does not apply and other distributions may be much more appropriate. For example, based on the experiments in our case study, we found that the Gamma distribution is a suitable candidate distribution. For the purpose of the our case study, we therefore introduce for each flex-run $f$ a Gamma distribution $\text{Gamma}(k, \theta)$ with mean $\mu_f$ and variance $\sigma_f^2$ to model the probability distribution of the load of the flex-run. Recall that the parameters $k$ and $\theta$ associated with the Gamma distribution are computed as follows:

$$
k = \mu_f^2 / \sigma_f^2, \quad \theta = \sigma_f^2 / \mu_f.
$$
We can determine whether a flex-run satisfies constraint (9) by inspecting whether the \( \alpha \)-th quantile of \( \text{Gamma}(k, \theta) \) exceeds the vehicle capacity \( \text{Cap} \). We remark that any other suitable probability distribution may be applied instead of the Gamma distribution.

Finding the minimum distance (or cost) \( c_f \) of a flex-run \( f \) corresponds to solving the well-known NP-hard Traveling Salesman Problem. Similarly, the ‘pricing problem’ in the delayed column generation approach that determines whether there exists a cost-improving flex-run contains the same underlying combinatorial problem. Therefore, if the potential number of locations that are served by a flex-run is large, this can limit the applicability of our approach. However, in many practical situations, the number of locations per vehicle is limited. As a first example, most vehicles (i.e., drivers) have a limited time ‘capacity’, and can only serve a maximum number of locations per flex-run. Furthermore, additional stops are often viewed as undesired, which is reflected in a quadratic (or higher polynomial) penalty function in the number of visits above a certain threshold. Finally, in certain cases management may simply impose an explicit hard limit on the number of visits per vehicle. In all these cases, the bound on the number of visits may very well yield a practically tractable (pricing) problem.

We note that it is conceptually straightforward to take into account additional restrictions when designing the flex-runs, such as time windows in which locations must be visited, precedence relations among locations, or stacking constraints, similar to what is done in column generation approaches for deterministic VRPs. Moreover, flex-runs naturally allow to model much more complex cost structures than simple distance accumulations. This is particularly important in industrial logistics applications that apply complicated cost patterns, as we will also see in our case study.

### 4.2 Set Covering Model

Let \( F \) be the set of all flex-runs for a given problem, and let the integer variable \( x_f \) represent the number of times that flex-run \( f \) is used in the solution. Our stochastic VRP with uncertain demands can now be formulated as the following chance-constrained set covering problem:

\[
\begin{align*}
\min & \sum_{f \in F} c_f x_f \\
\text{s.t.} & \sum_{f \in F} p_{fv} x_f = 1 \quad \forall v \in V \setminus \{v_0\} \\
& x_f \geq 0, \text{ integer} \quad \forall f \in F.
\end{align*}
\] (10)
In most practical cases, the set $F$ will be too large to be included entirely in the model. It is therefore natural to solve an approximate model that contains only a small subset of flex-runs. If required, a delayed column generation procedure can be applied to generate additional cost-improving flex-runs. We will see that in our case study, the latter part is not necessary, as we only generate flex-runs serving up to three customer locations.

4.3 Recourse Model

Inherently, an optimal solution to the chance constrained problem (10) may lead to overloaded vehicles when the uncertain demands are realized. In the practical setting that we envision, the flex-runs are fixed routes whose cost does often not depend on the weight that is carried. This is for example the case when a full trailer is hired. In those situations, the (monetary) cost of the existing flex-run does not change if we decide to have parts of the load be picked up by an additional ad-hoc vehicle. We therefore estimate the overload cost of a flex-run by the additional cost of shipping the expected overload amount of that flex-run. In particular, we will optimistically assume that the overload amount will be hauled first from the cheapest city, which corresponds to common practice in our case study. In other words, we propose as a recourse model to introduce one or more new flex-runs that are dedicated to pickup the overload amount, while utilizing the original flex-run to its maximum capacity.

The classical recourse action that is used in the literature to repair an overloaded route, is to follow the vehicle’s original route, send it to the depot when it reaches its capacity, and then proceed the route to pick up the remainder of load. Our proposed model deviates from the classical model in that the cost and structure of the original route remains unchanged, and that the overload amount is not combined in a route, but evaluated separately for each location on the route.

In order to compute the expected overload cost of a flex-run, we propose the following model. We represent the overload amount for a given flex-run $f$ as a random variable $o_f$, i.e., $o_f = \max(0, X_f - \text{Cap})$. The expected overload amount for $f$ can be computed by integrating $o_f$, which inherits the non-zero part of its distribution from $X_f$. For example, when $X_f$ assumes a Gamma distribution, the expected overload amount takes the following form:

$$
\mathbb{E}(o_f) = \int_{\text{Cap}}^\infty X_f^{k-1} \frac{e^{-X_f/\theta}}{\theta^k \Gamma(k)} dX_f,
$$

where $\Gamma(k)$ is the gamma function parameterized by $k$, i.e., $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t} dt$. 

As described above, the overload amount is distributed over the locations in an ‘optimistic’ way. We order the locations \( j \) visited by \( f \) by non-decreasing distance from the depot \( c_{jv_0} \), and we define the flex-run \( f[i] \) to be \( f \) restricted to the first \( i \) locations in this order. Furthermore, we define \( f[0] \) as the ‘empty’ flex-run, i.e., \( p_{f[0]} = 0 \). Thus, \( o_{f[i]} \) is the random variable representing the overload amount of the flex-run \( f[i] \). This allows us to formulate our expected overload cost for \( f \), denoted by \( c'_f \), as:

\[
c'_f = \sum_{i=1}^{\lvert f \rvert} c_{iv_0} \left( \mathbb{E}(o_{f[i]}) - \mathbb{E}(o_{f[i-1]}) \right).
\]  

We note that in most practical cases, \( c_{iv_0} \) will actually represent the unit shipping cost from location \( i \) to the depot rather than the distance in (11). To illustrate equation (11), consider a flex-run on three locations A, B, and C, such that \( c_{A0} \leq c_{B0} \leq c_{C0} \). We first consider the case that A overloads \( f \) by itself, with cost \( c_{A0} \). Then we consider the case that A and B together overload \( f \), but only pay for the marginal contribution of B relative to A; observe that since \( f[i] \) includes all pickups of \( f[i-1] \), we have that \( \mathbb{E}(o_{f[i]}) - \mathbb{E}(o_{f[i-1]}) > 0 \). Finally, we consider the case that A, B, and C together overload \( f \), but again we only pay for the marginal contribution of C relative to A and B. In this way, the total expected overload amount is distributed over the cheapest locations first.

The corresponding stochastic problem with recourse then corresponds to model (10) in which the objective function is changed into:

\[
\min \sum_{f \in F} (c'_f + c_f) x_f.
\]

Since a recourse action only affects the individual overloaded flex-run, the other parts of model (10) remain the same.

Observe that our proposed recourse model is not ‘optimal’ in that other models may provide a lower expected overload cost, for example by redistributing the load over two or more entirely new flex-runs. The purpose of our model is different: we wish to capture a common practical strategy in a generic mathematical model.

5 Case Study: Bosch/Siemens Home Appliances

We have applied our methodology to evaluate the incoming freight logistics network for the Bosch/Siemens Home Appliances Corporation (abbreviated as B/S/H/) located in New-
Bern, North Carolina. This site contains three manufacturing plants, dishwashing, cooking, and laundry that operate almost independently.

B/S/H/ has about 600 active suppliers for its three assembly plants in New Bern, NC. Among these, only 75 suppliers make up 80% of the total shipping volume. We chose to focus our study on these highly active suppliers. Each makes at least 40 shipments per year to one or more of the B/S/H/ plants in a weekly or bi-weekly schedule, making them practical targets for one or more weekly flex-runs.

Third-party logistics providers are hired externally, and most often ‘full-truckloads’ can be negotiated. For smaller shipments, ‘less-than-truckloads’ are usually applied. A full-truckload, or FTL, is a full, dedicated trailer with a typical weight capacity of 45,000 lb. In contrast, freight shipped by less-than-truckload, or LTL, is grouped with other shipments by the carrier and is subject to lower weight limits. The exact pricing of FTL and LTL shipments depends on many factors, the most important of which are the total distance traveled and total weight transported. In addition, the recently introduced fuel surcharges play an increasingly important role. However, FTL shipments are usually more efficient (in terms of the unit cost per pound per mile) when the amount shipped is at least 10,000 lb. For smaller amounts, LTL shipments typically provide a lower unit cost. The current practice is to order a single vehicle per order, based on the weight: LTL for orders under 10,000 lb, and FTL otherwise. For this reason, many orders between 10,000 and the maximum 45,000 lb are carried on partially-empty or even mostly-empty FTL trucks (see Table 1).

<table>
<thead>
<tr>
<th>Type</th>
<th>Amount</th>
<th>Cost</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTL</td>
<td>25%</td>
<td>37%</td>
<td>n/a</td>
</tr>
<tr>
<td>FTL</td>
<td>72%</td>
<td>47%</td>
<td>38%</td>
</tr>
<tr>
<td>Other</td>
<td>3%</td>
<td>16%</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 1: Shipping situation at B/S/H/ (2007 data). For each shipment type the relative amount shipped and relative cost are given.

Even though B/S/H/ uses FTL routes whenever possible, LTL routes are applied in many cases, especially when demand is greater than anticipated. Namely, when the amount to be shipped would cause an overload of the FTL truck, the remaining freight must be transported by an LTL shipment. The current shipping situation at B/S/H/ is depicted in Table 1. It shows the relative amount shipped and the relative cost spent for each of the
shipment types LTL, FTL, and ‘Other’, where ‘Other’ refers to ad-hoc shipping methods such as parcel shipment. Note that even though 25% of all freight is shipped using LTL, it constitutes 37% of the total transportation costs. FTL, on the other hand, is used to ship 72% of all freight, and only constitutes 47% of the total cost.

Market fluctuations have a huge impact on sourcing volumes. The current operating plan is able to react quickly to changes in volume: each supplier is ordered from independently, so a change in volume (either positive or negative) does not affect the shipments of other suppliers. Moreover, the addition of LTL trucks allows rapid adaptation to volume increases. This flexibility comes with a cost, however. LTL shipments are much more expensive relative to FTL shipments. Furthermore, the weekly or bi-weekly FTL routes are often under-utilized: the average FTL utilization is 38% (see Table 1). This means that, on average, FTL trucks are using only 38% of their capacity.

The purpose of our case study is twofold. First, we wish to identify the potential cost savings due to shipment consolidation. Second, we need to assess the impact of shipment consolidation to the volume fluctuations. In the remainder of this section, we will describe how we have implemented our chance-constrained model and stochastic model with recourse for this particular application.

5.1 Generating the Flex-Runs

We remark that for this case study, the vehicles are not stationed at the depot (B/S/H/), but instead commence their route at the first pickup location. We can mimic the classical VRP model by introducing auxiliary distances $c_{v_0i} = 0$ between the depot $v_0$ and each supplier location $i$.

In order to model splittable shipments for the suppliers, we apply a granularity of 2,000 lb. That is, a vehicle can pick up a positive integer multiple of 2,000 lb at each supplier. We apply a suitable upper bound on the integer multiplier following from historical data for each supplier, i.e., the maximum weekly shipment.

We first generated one-stop LTL routes for all suppliers. If the maximum weekly shipment (based on historic data) from a supplier is less than 10,000 lb, we decided to add only a single LTL route for the entire shipment to avoid an excessive inconvenience on the supplier. Otherwise, we generate an LTL route for each integer multiple of 2,000 lb not exceeding 10,000 lb (the cut-off point from LTL to FTL shipment).
We next generated FTL routes for all combinations of one, two and three suppliers. The limit of at most three suppliers was motivated by the technical restriction of computer memory, but also B/S/H/ management indicated that more stops are undesirable from a coordination point of view. For 75 suppliers, the number of all such combinations is \( \sum_{k=1}^{3} \binom{75}{k} = 70,375 \). For each combination, we checked all possible orders in which the suppliers could be visited before driving to B/S/H/, and selected the sequence with the lowest total mileage. We used Microsoft MapPoint NA 2006 with the MileCharter add-in to compute a distance matrix for this calculation. We need several copies of each FTL route to cover all meaningful combinations of pickup proportions. We enumerated all these combinations, and kept only those for which the chance constraint (8) on the vehicle capacity was not violated.

We next discuss how we compute the cost of each flex-run. The pricing of the LTL flex-runs is relatively straightforward, as they are simply based on the distance traveled and the shipment load. For each of the generated LTL routes, we looked up a rate quote on the web site of the LTL carrier used by B/S/H/ in the region of the supplier.

For FTL flex-runs, pricing is much more involved. Trucks ordered from FTL carriers are priced according to a fixed, negotiated rate per load. Each year, B/S/H/ distributes a list of locations to each of its carriers from which it plans to order many FTL shipments. A multiple-round bidding process takes place that identifies the lowest price a carrier will accept for each shipment made from each location. An exclusive contract is granted to each carrier for those locations on which it is the lowest bidder. B/S/H/ can then order FTL shipments at these fixed prices when needed. Because FTL shipments are not priced by weight, they are somewhat flexible to changes in quantities if they are under capacity.

For our model, we decided to approximate the FTL pricing structure as follows. We selected a representative subset of existing negotiated FTL routes used by B/S/H/. In Figure 1 we plot the cost of those routes against their length. From this figure, it is apparent that a close relationship exists between the mileage of a route and the minimum price an FTL carrier would accept to haul it. We fit a quadratic curve to this data and used it to estimate prices for our proposed FTL routes. We also added a stopover charge found on the carrier’s website for those routes visiting more than one supplier.

For the recourse model, we determine the expected overload cost for each flex-run using the formula (11). We have used MATLAB to evaluate the expected overload amounts for the respective flex-runs \( f[i], i = 1, 2, 3 \).
Figure 1: Cost (in USD) of previously negotiated FTL routes with respect to their length (in miles). The cost numbers are rescaled for confidentiality.

The resulting set of flex-runs, containing all LTLs and FTLs of one to three suppliers using all allowed weight combinations, number around 600,000. Consequently, our integer programming model consists of about 600,000 variables and 75 constraints (for 75 suppliers).

5.2 Computational Results

In this section we provide a detailed analysis of the solution we proposed to B/S/H/. In addition to a single solution, B/S/H/ management wanted us to test the robustness of our model parameters. These parameters include the maximum number of stops on each flex-run, the minimum service level $\alpha$ that we require for each flex-run, and the shape of the weight distribution assumed for each flex-run load. We performed several experiments to test different values of the parameters and confirm the accuracy of the model.

Data was collected from records kept by a third-party logistics provider. The data identified the date, source, type, weight, and cost of each shipment. Unfortunately, the records did not indicate to which of the three plants each shipment was made. Thus, while we could analyze the benefits of consolidating any set of supplier shipments, we could not analyze the intermediate, and possibly more manageable, benefit of consolidating only within each plant. Throughout this section, cost figures have been rescaled by a fixed multiplier for confidentiality.

For our experiments we have used the integer programming solver CPLEX 11 [10] on
a 2.4Ghz machine with 8GB of RAM running Windows Vista 64 Pro. For the 75-supplier instances, we report the solutions after a duality gap of 1% was reached, i.e., the solution found is provably within 1% of the optimal solution (to our model). Finally, we note that for the larger problem instances on around 600,000 variables, most of the available RAM was used. All reported solutions were found within 24 hours total.

5.2.1 Number of Stops per Route

Our truck routing model contains 75 supplier locations that must be visited by one or more trucks each week. As described in our route generation procedure, we considered only those routes visiting at most three suppliers. This both limits the size of the model for computational tractability and limits the amount of coordination required for B/S/H/ to implement the solution. To study the effects of this maximum-stops parameter on an optimal routing cost, we tested the inclusion of routes with more stops on a smaller example problem containing 15 suppliers. For a problem of this size, the number of these routes is small enough to enumerate and solve to optimality. We also used a distribution of weight and locations similar to our B/S/H/ data.

<table>
<thead>
<tr>
<th># Suppliers</th>
<th>Max Stops per Route</th>
<th>Min Reliability</th>
<th>Optimal Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1</td>
<td>95%</td>
<td>$30,553</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>95%</td>
<td>$19,292</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>95%</td>
<td>$14,427</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>95%</td>
<td>$12,916</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>95%</td>
<td>$12,714</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>95%</td>
<td>$12,714</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>95%</td>
<td>$12,714</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>95%</td>
<td>$12,714</td>
</tr>
</tbody>
</table>

Figure 2: The effect of increasing the number of stops per route on the optimal shipping cost, for a 15-supplier benchmark problem.

In Figure 2 we display the results of this experiment by means of a table and a plot. In this figure, ‘# Suppliers’ indicates the number of suppliers in the problem, ‘Max Stops per Route’ indicates the maximum number of stops (i.e., suppliers) on each route, ‘Min Reliability’ indicates the reliability of each route to be within the truck capacity (in this experiment it is set to 95%), and ‘Optimal Cost’ indicates the optimal weekly shipping cost for this problem.
## Table 2: The effect of increasing the number of stops per route on the shipping cost and expected savings, for the 75-supplier problem of B/S/H/.

As the results show, cost savings improve as more stops are added to the feasible routes, but the marginal value of the extra stops diminishes quickly. To observe this effect on our full 75-supplier problem, we tested our model with a maximum of one, two, and three stops per route. The results are depicted in Table 2. In this table, we compare four different transportation schedules: the original schedule of B/S/H/ and our three optimized schedules based on routes including a maximum of one, two, and three stops per route, respectively. Here ‘# Feasible Routes’ indicates the number of feasible routes that were generated, ‘Est. Weekly Fixed Cost’ indicates the estimated weekly fixed transportation cost of the schedule, and ‘Savings’ indicates the savings of the optimized schedules with respect to the original schedule.

Note that the optimized schedule with 1 stop per route has a lower cost than the original schedule. This is the result of multiple LTL routes being consolidated into single FTL routes for suppliers shipping to multiple plants at B/S/H/.

### 5.2.2 Reliability of Remaining within Truck Capacity

The significant cost reduction of our solution is attributable to an increased use of FTL routes, since FTL costs less per pound of capacity than LTL. But to exploit this cost reduction, one must consistently have enough freight to utilize the large, fixed capacity of the FTL. This can be difficult to achieve under conditions of high demand variability: while assigning several suppliers to an FTL route may reduce per-unit costs, it may also result in frequent overloading of the truck capacity. In such case, an additional LTL route may be necessary, negating the savings of the FTL route and increasing managerial costs.

To balance the two goals of minimizing cost and maintaining reliability, we next consider the effect of changing the reliability level $\alpha$. In Table 3 we report the results of an experiment

<table>
<thead>
<tr>
<th>Schedule</th>
<th># Suppliers</th>
<th>Max Stops per Route</th>
<th>Min Reliability</th>
<th># Feasible Routes</th>
<th>Est. Weekly Fixed Cost</th>
<th>Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>75</td>
<td>1</td>
<td>-</td>
<td>75</td>
<td>$134,063</td>
<td>-</td>
</tr>
<tr>
<td>Optimization 1</td>
<td>75</td>
<td>1</td>
<td>95%</td>
<td>293</td>
<td>$123,331</td>
<td>8.0%</td>
</tr>
<tr>
<td>Optimization 2</td>
<td>75</td>
<td>2</td>
<td>95%</td>
<td>13,052</td>
<td>$105,337</td>
<td>21.4%</td>
</tr>
<tr>
<td>Optimization 3</td>
<td>75</td>
<td>3</td>
<td>95%</td>
<td>592,975</td>
<td>$100,000</td>
<td>25.4%</td>
</tr>
</tbody>
</table>
Table 3: The effect of increasing the reliability for each route to remain within capacity on the shipping cost and expected savings, for the 75-supplier problem of B/S/H/.

<table>
<thead>
<tr>
<th>Schedule</th>
<th># Suppliers</th>
<th>Max Stops per Route</th>
<th>Min Reliability</th>
<th># Feasible Routes</th>
<th>Est. Weekly Fixed Cost</th>
<th>Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>75</td>
<td>1</td>
<td>-</td>
<td>75</td>
<td>$134,063</td>
<td>-</td>
</tr>
<tr>
<td>Optimization 3a</td>
<td>75</td>
<td>3</td>
<td>99%</td>
<td>291,110</td>
<td>$121,899</td>
<td>9.1%</td>
</tr>
<tr>
<td>Optimization 3b</td>
<td>75</td>
<td>3</td>
<td>95%</td>
<td>592,975</td>
<td>$100,000</td>
<td>25.4%</td>
</tr>
<tr>
<td>Optimization 3c</td>
<td>75</td>
<td>3</td>
<td>90%</td>
<td>851,446</td>
<td>$87,518</td>
<td>34.7%</td>
</tr>
</tbody>
</table>

Table 4: Estimating the overload cost for the complete set of generated routes, and for the routes that are included in the optimal solution. These results are based on the 2007 data (12 months) from B/S/H/.

<table>
<thead>
<tr>
<th>All Feasible Routes</th>
<th>Min Fixed Cost Routes</th>
<th>Min Total Cost Routes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximate Weekly Fixed Cost</td>
<td>$2,874,994,855</td>
<td>$100,000</td>
</tr>
<tr>
<td>Expected Weekly Overload Cost</td>
<td>+ $19,177,002</td>
<td>+ $4,091</td>
</tr>
<tr>
<td>Total Weekly Cost</td>
<td>= $2,894,171,856</td>
<td>= $104,090</td>
</tr>
<tr>
<td>Actual Overload Cost</td>
<td>$11,063,102</td>
<td>$1,821</td>
</tr>
</tbody>
</table>

5.2.3 Estimating the Overload Costs

In order to further quantify the cost of unreliability in our solution, we wished to measure the expected cost incurred during the 5% or less of cases in which FTL route capacity is overloaded. For this, we apply our recourse model.

Table 4 shows the expected overload cost associated with our proposed solution. We estimate this cost for all feasible routes that were included in the model (‘All Feasible Routes’), and for the subset of routes included in the solution to two optimizations. The first (‘Min Fixed Cost Routes’) corresponds to the optimal solution that we have considered thus far, using a fixed cost for each route. The second (‘Min Total Cost Routes’) adds to the fixed cost the costs associated with the expected variable costs due to overloads.
cost the overload cost for each route. Therefore, the objective is slightly higher, and the solution changes slightly.

For each of these three route sets, we report the approximate weekly fixed cost of the routes, the expected weekly overload cost, and the total weekly cost of the routes. We compare this (in the last row of Table 4) to the actual overload costs, based on the B/S/H/ routes of 2007. Note that the latter is computed from the data itself, not the distribution fit to the data.

A first observation is that the expected overload cost for the routes included in the optimal solution is relatively high, around 4% of the weekly cost, compared to that of all feasible routes, which is around 0.67%. From Table 4 we further derive that the overload costs are somewhat over-estimated, but at the same time constitute only a marginal amount of the overall transportation costs. Further, adding the overload costs to the objective did only slightly change the optimal solution, and we therefore omitted the overload costs in our recommended solution.

5.2.4 Analyzing the Proposed Solution

As stated earlier, we chose to recommend to B/S/H/ the optimal solution using the 95% reliability level with respect to the overload risk. We next provide a detailed analysis of this solution and compare it to the original shipping situation at B/S/H/.

We first discuss the structure of the original routes and the optimized routes in terms of the amount of LTL and FTL shipments, presented in Table 5. To allow a meaningful comparison, the reported costs of the original routes are based on the model we presented in section 5.1. In Table 5, we report for each route type the average mileage, the average cost per week, the average ton-mileage per week, and the average cost per ton-mile. A first observation is that the number of routes dramatically decreased from 165 to 60 routes in total, most of which are 3-stop FTLs. Interestingly, the average mileage for the resulting optimized LTL routes is higher than the original LTL routes. Indeed, an analysis of our solution showed that the best candidates for optimized LTL routes are those that are geographically isolated from the others, while other factors are low shipping weight and high variability.

To identify which routes (and suppliers) attribute most to the total cost savings of about 25%, we present in Table 6 the average unit costs for the different route conversion: from LTL to Optimized LTL, from LTL to Optimized FTL, and from FTL to Optimized FTL. The most important contribution stems from the conversion of LTL routes into Optimized FTL
Table 5: The structure of the original and optimized routes.

<table>
<thead>
<tr>
<th>Route Type</th>
<th>#/week</th>
<th>Avg. Mileage</th>
<th>/week</th>
<th>ton-mi/week</th>
<th>/ton-mi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Stop LTL</td>
<td>126</td>
<td>623 mi</td>
<td>$62,853</td>
<td>122,706</td>
<td>$0.49</td>
</tr>
<tr>
<td>1-Stop FTL</td>
<td>39</td>
<td>436 mi</td>
<td>$71,210</td>
<td>258,966</td>
<td>$0.29</td>
</tr>
</tbody>
</table>

Optimized Routes, 95% Reliability Level

<table>
<thead>
<tr>
<th>Route Type</th>
<th>#/week</th>
<th>Avg. Mileage</th>
<th>/week</th>
<th>ton-mi/week</th>
<th>/ton-mi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Stop LTL</td>
<td>19</td>
<td>880 mi</td>
<td>$11,873</td>
<td>30,813</td>
<td>$0.39</td>
</tr>
<tr>
<td>1-Stop FTL</td>
<td>9</td>
<td>128 mi</td>
<td>$7,871</td>
<td>23,612</td>
<td>$0.33</td>
</tr>
<tr>
<td>2-Stop FTL</td>
<td>9</td>
<td>305 mi</td>
<td>$11,807</td>
<td>49,309</td>
<td>$0.24</td>
</tr>
<tr>
<td>3-Stop FTL</td>
<td>23</td>
<td>897 mi</td>
<td>$68,449</td>
<td>341,852</td>
<td>$0.20</td>
</tr>
</tbody>
</table>

Table 6: Cost savings appear when converting the original routes into optimized routes.

<table>
<thead>
<tr>
<th>Type Conversion</th>
<th>Original Routes</th>
<th>Optimized Routes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$/ton-mi</td>
<td>95% reliability</td>
</tr>
<tr>
<td>LTL → LTL Optimized</td>
<td>$0.44</td>
<td>30,813</td>
</tr>
<tr>
<td>LTL → FTL Optimized</td>
<td>$0.44</td>
<td>91,893</td>
</tr>
<tr>
<td>FTL → FTL Optimized</td>
<td>$0.29</td>
<td>258,966</td>
</tr>
</tbody>
</table>

routes (from $0.44 to $0.21 per ton-mile), while converting LTLs and FTLs into optimized LTLs and FTLs contribute relatively less to the savings.

Finally, we analyze the truck utilization of our optimized solution, as compared to the original shipping situation at B/S/H/. As was reported in Table 1, the current shipping situation uses only 38% of the FTL truck capacity, on average. We report in Table 7 the average truck utilization of our proposed solution, as well as the ‘peak utilization’. The latter is defined as the average of the highest loads over all trucks. For example, the 10% peak utilization collects for all trucks the 10% highest load volumes, and the average is reported as the 10%-peak volume (in percentage of the capacity). Similarly for the 5%-peak volume. The results in Table 7 clearly show an improvement in average truck utilization with respect to the original situation (from 38.1% to 48.5%). We note that the average truck utilization is deliberately kept at a relatively low level by our optimization model, to allow a reliability
buffer for the flex-runs. That this is indeed needed is shown by the peak utilization of the trucks. For example, for the 95% reliability level, the 10%-peak utilization is 81.9% of the truck capacity. This means that many flex-runs will in fact be almost completely filled during roughly 10% of the schedule execution. We note that for the 90% reliability level, the higher chance of overloading is witnessed by the 5%-peak utilization of 115.6% of the truck capacity (i.e., an overload of 15.6%).

5.3 Implementation in Practice

From the perspective of B/S/H/, the initial goal of this project was to gain insight in the potential cost savings for their inbound freight logistics operations. If these cost savings are high, a company-wide implementation could be worthwhile. B/S/H/ further envisioned that the insight gained by this project (particularly the cost savings) could be useful during the contract negotiations with their logistics providers.

We provided B/S/H/ with the findings reported in this work. In particular, we presented them the optimal solution with a 95% reliability level for the overload risk. Our reported expected cost savings of up to 25% led B/S/H/ to initiate an implementation of our solution, first on a small scale.

Based on the solution we reported, B/S/H/ selected a subset of most profitable flex-runs and locations, as candidates to implement in a first phase. For this subset, B/S/H/ provided more detailed information and restrictions that needed to be taken into account. For example, for some of the routes stacking and volume restrictions were added, while for other routes it was important to respect a certain order to visit the suppliers. Our model can easily handle all these additional restrictions during the route generation process: we only include those routes that are feasible with respect to all restrictions. From the solution that was obtained by re-optimizing this model, B/S/H/ selected five flex-runs for actual

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>95% reliability</th>
<th>90% reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Utilization</td>
<td>38.1%</td>
<td>48.5%</td>
<td>55.5%</td>
</tr>
<tr>
<td>10%-Peak Utilization</td>
<td>-</td>
<td>81.9%</td>
<td>96.5%</td>
</tr>
<tr>
<td>5%-Peak Utilization</td>
<td>-</td>
<td>96.6%</td>
<td>115.6%</td>
</tr>
</tbody>
</table>

Table 7: Comparing the average and peak FTL truck utilization of the optimized solution with the original shipping situation.
implementation in practice.

It should be noted that our proposed solution can have major implications throughout the organization at B/S/H/. Most importantly, it requires communication between different logistics planners, sometimes even between different plants. It could therefore be useful in practice to limit such communication by introducing certain planning protocols for shared trucks. Alternatively, one could try to make the communication more transparent through a web-based plant-wide planning tool. In any case, this is an important issue that must not be ignored during the implementation.

Another important issue is the reward and incentive structure at manufacturing companies such as B/S/H/. Currently at B/S/H/, logistics planners are mainly evaluated based on inventory levels (the lower the better), while transportation costs are mostly ignored. Since our proposed solution focuses on optimizing the transportation costs, there is no immediate incentive for logistics planners to put extra effort in the successful implementation of our solution. It would therefore be advisable if the evaluation of the logistics planners would take transportation costs into account as well.

6 Summary and Conclusions

We have introduced so-called flex-runs to model flexible milk-runs for stochastic vehicle routing problems (VRPs) with uncertain splittable demands. These flex-runs have been applied in the well-known set covering formulation of the problem. The main benefit of flex-runs in the context of stochastic VRPs is that the interdependencies between the demand fluctuations at the pickup locations can be exploited to design more robust consolidated shipment routes. We have also introduced a new recourse model, based on the practical strategy of hauling the excess shipment amount with ad-hoc single-stop vehicles.

We have performed a case study for the manufacturing site of Bosch/Siemens Home Appliances in North America, with the purpose of evaluating their incoming freight logistics network. To this end, we have designed an optimized set of flex-runs based on real shipment data, combining up to three suppliers per flex-run. Our most important finding is that with our solution, expected cost savings of up to 25% can be achieved with respect to the current shipping situation, while at the same time, the robustness of the schedule with respect to changes in shipping volume increases. An initial implementation of our proposed solution
is currently initiated at Bosch/Siemens. Our approach is not restricted to the situation at Bosch/Siemens however, but in fact is broadly applicable.

References


