Locomotive Planning at Norfolk Southern: 
An Optimizing-Simulator using Approximate Dynamic Programming

Warren B. Powell  
powell@princeton.edu  
(609)273-0218

Belgacem Bouzaiene-Ayari  
belgacem@princeton.edu

Department of Operations Research and Financial Engineering  
Sherrerd Hall  
Princeton University  
Princeton, NJ 08544

Coleman Lawrence  
coleman.lawrence@nscorp.com

Clark Cheng  
clark.cheng@nscorp.com

Sourav Das  
sourav.das@nscorp.com

Ricardo Fiorillo  
ricardo.fiorillo@nscorp.com

Norfolk Southern Corporation  
1200 Peachtree St. NE, Box 12-117  
Atlanta GA 30309

July, 2013
Abstract

For decades locomotive planning has been approached using the classical tools of mathematical programming, resulting in very large-scale integer programming models which are simultaneously beyond the capabilities of modern solvers, while still requiring a host of simplifying assumptions that limit the use of the models for the analysis of important planning problems. The primary interest of Norfolk Southern was a model that could assist with fleet sizing, and yet the cumulative effect of simplifications required to produce a practical integer programming formulation produced a model unable to accurately predict fleet size. We use the modeling and algorithmic framework of approximate dynamic programming (ADP), which uses an intuitive balance of simulation and optimization with feedback learning, to produce a highly detailed model that accurately calibrated against historical metrics. The result was a model that could be used to plan fleet size and mix, as well as being sensitive to a wide range of operating parameters. At the same time, the intelligence of ADP allows the model to adapt to a wide range of scenarios.
Locomotive planning is one of the most complex operational problems in freight transportation. Planners have to take into consideration a host of operational characteristics that describe locomotive operations to best utilize the fleet to meet the service requirements of the trains. Locomotive fleets can represent billions of dollars in investments, and as a result railroads have every incentive to manage this investment as efficiently as possible. The complexity of the problem has put it well past the capabilities of even today’s advanced optimization solvers. Completely overlooked in these models are the important sources of uncertainty such as transit time delays, the dynamics of scheduling commodities such as coal and grain, and the ever present problem of equipment failures and maintenance.  

A large railroad may have billions invested in their fleet of locomotives. Too many locomotives mean that hundreds of millions are invested in equipment that is not yielding a return. However, a failure to maintain a sufficiently large fleet translates to delayed trains and service failures that can seriously impact revenue. Despite the scale of this investment, it is not unusual for a large railroad to plan their locomotive fleets using a simple linear regression that relates operating statistics such as forecasted tonnage and operating speeds to the amount of power required to run the railroad.

Given the size of the investment, railroads have tried for decades to tap the power of optimization tools to manage their fleets more effectively. Booler (1980) presents a very early attempt at using linear programming to solve a scheduling model. Chih (1986) and Chih et al. (1990) describe an implementation of an integer programming model for locomotive scheduling at the Burlington Northern Sante Fe Railroad (see also Forbes (1990)). This early work struggled with the limitation of integer programming algorithms, despite using highly simplified models of locomotive operations. A number of advances have been made in the design of specialized algorithms to solve integer programming formulations locomotive scheduling. Ziarati (1999) presents a new branch and cut algorithm. Ahuja et al. (2005) describe a heuristic based on very large scale neighborhoods to find near-optimal schedules for locomotives which considers consist breakups and the desire for weekly patterns in the flows of locomotives. Vaidyanathan et al. (2008a) provides a detailed model of the locomotive routing problem capturing a number of operational constraints with an adaptation of their large neighborhood search strategy (see also Vaidyanathan et al. (2008b) for additional experimental work). Cordeau et al. (2000, 2001) describe the use of Benders decomposition for the simultaneous assignment of cars and locomotives. At the same time, it is important to recognize major advances to general purpose integer programming solvers such as Cplex and Gurobi that have occurred since 2005. Below, we report on experiments with Cplex 12 running on a single machine with 1 terabyte of RAM and 64 threads.

This paper describes a multiyear effort to develop a family of locomotive planning models for Norfolk Southern Railroad. The result is the Princeton Locomotive And Shop MAnagement system (PLASMA), which has been implemented at Norfolk Southern for strategic planning and short term operational planning. PLASMA has been imbedded in a larger information system developed at Norfolk Southern called the Locomotive Assignment and Routing System (LARS). As of this writing, the system has been used for strategic fleet sizing for several years and has become an integral part of the company’s network and resource planning processes. The operational forecasting system has been undergoing

---

1 This paper is based heavily on Powell et al. 2012.
extensive user acceptance testing while Norfolk Southern has been upgrading its information systems to improve the accuracy of some of the data.

**Locomotive operations**

Locomotives are described by a host of attributes including horsepower and tractive effort, the owning railroad, its maintenance status (e.g. days until the next federally mandated maintenance appointment), and equipment details such as communications gear (for coordinating multiple locomotives). Locomotives typically need to be bundled together into consists of two to four locomotives which are needed to pull a particular train. The process of connecting locomotives is time-consuming, requiring the connecting and testing of cables that allow the set of locomotives to work as a single unit. Trains often arrive from a neighboring railroad using locomotives owned by that railroad (known as “foreign power”). Normally foreign locomotives are returned to the owning railroad, but it is possible to use these locomotives, but only within negotiated limits. If the locomotive is returned, this has to be handled through pre-defined exchange points.

The trains themselves also have numerous attributes. The number of locomotives needed to pull a train depends on the weight of the train, the speed requirements (merchandise trains need to move more quickly than coal and grain trains), and the steepest grade that the train has to navigate. Trains have different service priorities, and a scheduler has to consider if there are not enough locomotives to move all the trains on time.

Shop routing (getting the locomotive to a maintenance shop) is one of the most complex issues facing a locomotive manager. Locomotives require government-mandated regular maintenance inspections on a periodic schedule (typically every 92 days). If a locomotive does not make its shop appointment on time, it has to be turned off and “towed” to the shop. Of course, on some trains it is possible to simply add an extra locomotive move it to its shop appointment, but it is better to use the locomotive productively. At the same time, the scheduler has to balance getting the locomotive to shop early (which costs productivity) or risk that it may arrive late (for example, by missing a critical connection). On top of this, there may be multiple shop locations that can service a locomotive. It is necessary to anticipate the number of locomotives that are scheduled at each shop in order to balance the loads and maintain a steady flow of work.

A separate issue that is widely discussed but rarely solved concerns the different types of uncertainty that plague locomotive operations. These include:

- **Transit time delays** – These can be as long as six to 12 hours for the shorter movements of an Eastern railroad, to more than a day for the long movements of the western railroads.
- **Dynamic schedule changes** – Planners also have to deal with the pattern of scheduling additional trains for commodities such as coal and grain, with as little as one or two days advance notice.
- **Shop delays** – Maintenance managers will provide estimates of when a locomotive will be ready to leave a shop, but these are just estimates and frequent calls to a maintenance facility are often needed to determine if a locomotive will be ready for a particular train.
• Equipment failures – Locomotives may fail unexpectedly, and this represents an additional source of uncertainty.

The model presented in this paper is designed to handle uncertainty, but production applications of the model have yet to exploit this capability.

**Deterministic optimization models**

The most common strategy for modeling locomotives as an optimization problem is to use the framework of multicommodity flows over a time-space network, depicted in figure 1. This figure shows space (vertically) and time (horizontally), where each node represents a terminal (in space-time), and the lines represent movements of locomotives from one terminal to another. The figure depicts three types of locomotives in different shades of gray. This classical model has to be modified to handle issues such as the coupling/uncoupling of locomotives to move a train, and routing to shop. However, perhaps one of the most difficult challenges is the modeling of train delays. A standard strategy is to replicate the movement of a train at different points in time, and then introduce constraints so that only one copy of the train is actually moved. This requires forcing the train to move at discrete points in time such as 2pm, 4pm and 6pm, when in fact the train may have to be delayed 37 minutes to handle a late-arriving inbound train. We found that the rounding of departure times to the nearest two hours (for example) produced scheduling errors that were unacceptable to the railroad.

![Image](image_url)

**Figure 1- Illustration of multicommodity network flow problem over a time-space network.**

Deterministic models are extremely hard to solve over long planning horizons, especially when modeling the ability to delay trains. Figure 2 shows the run times as the horizon is extended for the Norfolk Southern fleet, using the same stopping tolerances for Cplex. Note the exceptionally fast execution time for the single-day horizon. For the smallest problem, locomotives are being assigned to at most a single
train. As the horizon grows, we have to model the cascading of train delays. With a horizon of as little as four days, the run times are already exceeding 50 hours.

**Model based on approximate dynamic programming**

Approximate dynamic programming is a modeling and algorithmic strategy that decomposes decisions over time (see Powell (2011) for an introduction to the concepts in this paper; approximate dynamic programming is closely related to the field of reinforcement learning, see Sutton and Barto (1998)). It was originally developed to handle problems which involve uncertainty. Our own work, however, has focused on its ability to decompose large deterministic problems, overcoming the dramatic increase in CPU times documented in figure 2. In fact, all of the production applications of ADP at Norfolk Southern have been conducted using a deterministic model.

Mathematically, we can write this problem in the form of a dynamic program. Let $R_t = (R_\alpha)_{a \in A}$ where $R_\alpha$ be the number of locomotives with attribute $a$ at time $t$, where $a$ captures the type of locomotive, its location, maintenance status (e.g. next time it is due in shop), special equipment (e.g. flush toilets and communication), and whether it is attached to other locomotives in a consist. Also let $D_t$ be the set of trains waiting to be moved (the “demand”). In the language of dynamic programming, our state variable at time $t$ is given by $S_t = (R_t, D_t)$. If $V_t(S_t)$ is the value of being in state $t$, then we can use Bellman’s optimality equation to write
\[
V_t(S_t) = \min_{x_t \in X_t} \left( \sum_{a \in A} \sum_{d \in D} c_{ad} x_{tad} + V_{t+1}(S_{t+1}(S_t, x_t)) \right)
\]

where \( x_{tad} \) is the number of locomotives with attribute \( a \) that we assign to train \( d \) in the set \( D_t \), and \( c_{ad} \) is the cost of this assignment (costs have to consider the need to break apart consists, as well as the appropriateness of a particular type of locomotive for a particular type of train). \( S_{t+1}(S_t, x_t) \) is the state of the system resulting from the decision vector \( x_t \).

A detailed mathematical formulation of the ADP model and algorithm is given in Bouzaiene-Ayari et al. [2013]. At the heart of the model is the locomotive assignment subproblem where we assign locomotives to trains at a particular point in time. This subproblem is depicted in figure 3, and it consists of two components: the assignment of locomotives to at most one train departing within a four-hour horizon, and piecewise linear “value function approximations” that capture the downstream value of locomotives in the future. Mathematically, we write the subproblem at time \( t \) as

\[
X_t^* (S_t) = \arg \min_{x_t \in X_t} \left( \sum_{a \in A} \sum_{d \in D} c_{ad} x_{tad} + V_{t+1}^{n-1}(S_{t+1}(S_t, x_t)) \right)
\]

This is the problem we solve at iteration \( n \), time \( t \), using the value function approximation \( V_{t+1}^{n-1}(S_{t+1}) \) from the previous iteration. If \( x_t^n \) is the optimal solution, we obtain the next state using

\[ S_{t+1}^n = S_{t+1}(S_t^n, x_t^n) \]

which is obtained by simulating from state \( S_t^n \) at time \( t \) to state \( S_{t+1}^n \) at time \( t+1 \), during iteration \( n \).

We approximate the value function with the separable, piecewise linear approximations that are depicted in figure 3. When we do this, we can solve the minimization problem in equation (1) as an integer program, which can be solved by commercial solvers in a few seconds, even when optimizing 2,000 locomotives for the entire railroad. We can use this integer program to estimate the marginal value of each locomotive, and these marginal values are used to estimate the piecewise linear functions.

The assignment model, which assigns individual locomotives to individual trains, is distinguished by its ability to handle a very high level of detail. For example, we can capture the precise arrival times of locomotives and desired departure time of the train (which may be delayed); we can model consist breakup and formation, routing of foreign power, routing of power toward shop locations, and the need for specific equipment types within a consist. We can model the fact that two locomotives are available at 12:42pm, while a third locomotive will not arrive until 1:37pm for a train that should have departed at 12:50pm, but which will have to be delayed until the last locomotive has arrived.

The value function approximations are the distinguishing feature of approximate dynamic programming. They are learned internally by the model by repeatedly simulating the process of assigning locomotives to trains, as shown in figure 4.
In figure 4, the gray boxes and dark black lines represent a single locomotive-to-train assignment problem with value functions, which we have overlaid on top of a space-time graph depicting the full set of flows over time. The figure depicts the assignment problem being solved at times t, t+1 and t+2 as it simulates through the full horizon of the model (in our work, the model steps forward in 4-hour time steps, although all assignments of locomotives to trains are made down to the minute). The bold arcs represent locomotives moving from one location to another, connecting with downstream value functions in the future. The simulation horizon is typically a month for strategic planning problems, but it may be a week for short-term operational planning. At each point in time, we stop and estimate the value of one more locomotive of each type, at each location. These marginal estimates are then smoothed into piecewise linear value functions at two levels of aggregation for each yard. Over time, these value functions capture the value of locomotives over an extended horizon. We use these value functions in the strategic planning model to determine fleet size and mix, since the value functions at the beginning of the horizon capture the marginal value of each type of locomotive over the entire horizon.

The process of stepping forward through time, solving sequences of relatively small problems, is a reason why approximate dynamic programming is often referred to as an “optimizing simulator.” While the integer program in Figure 3 still has thousands of integer variables, it can be solved using commercial packages such as Cplex in just a few seconds. The value functions allow us to produce solutions where a decision now anticipates the future. For example, a train may only require two locomotives, but we may assign three or four locomotives if we need to reposition power to a location that needs extra capacity. We can also think about the value of different types of locomotives in the future. The steps of the approximate dynamic programming algorithm are described in figure 5.
The methodology easily adapts to handle uncertainty in transit times, train schedules and locomotive failures. As we simulate decisions forward in time, we can use randomly sampled variables from historical distributions. Laboratory research (Bouzaiene-Ayari et al. [2013]) has shown that this produces very robust dispatch policies. However, Norfolk Southern has not as yet adopted the use of policies that were trained using such a sampling policy.

![Figure 4-Illustration of ADP policy stepping forward through time, solving sequences of small locomotive assignment problems with value function approximations.](image)

**Model calibration and validation**

The model went through several years of careful calibration against historical performance. This required the painstaking examination of detailed assignments, along with the comparison of high level performance metrics. The process involved the iterative identification and correction of data errors, as well as enhancements in the model and, from time to time, improvements in the basic algorithm. For example, it was through this process that we identified the need to use two layers of aggregation in the value function approximations as shown in figure 3. The most common data problems arose in the initial location of locomotives, and the representation of the train schedule and tonnage requirements. Examples of modeling problems included changes required in the handling of foreign power and the rules for consist formation.

In addition to the careful examination of individual assignments, Norfolk Southern focused on train delay as the most important metric of overall performance. Matching train delay at a system level is an extremely difficult target because it requires that the model match locomotive productivity almost perfectly. For example, it is important that we accurately capture the costs and time required for breaking up locomotive consists. If we ignore this component, we would over-represent the ability to use power to move trains, which in turn would underestimate train delay.
In the early stages of the calibration process, the model would produce delays that were an order of magnitude larger than history, largely as a result of data errors that had locomotives hopelessly out of position. It is not possible to match historical performance simply by tuning parameters within the model. It was essential that the detailed assignments pass the examination of experienced schedulers. This process was simplified by a powerful diagnostic tool called Pilotview (figure 6) that we developed for complex resource allocation problems such as this.

We proceeded by creating a curve from estimates of total train delay as a function of the fleet size. After finally getting the model to closely match historical performance, we repeated the exercise with an entirely new dataset. The result is the curve shown in figure 7, which shows a very close match between the curve and the historical delay at the current effective fleet size. Also note that the relationship between fleet size and train delay is smooth and predictable. Achieving this behavior with a model that captures this level of detail is actually quite hard, as it requires that we have the ability to model train delays continuously. It also suggests that the behavior produced by the value function approximations is quite smooth, a result that we did not anticipate at the start of the project.

Figure 5-Flow chart of steps of the approximate dynamic programming algorithm.
Other forms of model validation involved testing the sensitivity of the model to key input parameters. One such test evaluated the effect of increasing the consist breakup cost to determine its impact on both the number of consists being broken and overall solution quality. Figure 8 shows the effect of increasing the consist breakup cost, using the rate of consist breakups with a cost of zero as a baseline. The chart demonstrates that increasing the consist breakup cost produces a steady decline in the number of broken consists. We note that this was achieved with no discernible reduction in the model objective function which captures repositioning costs and penalties for delayed trains.

Figure 6 - Snapshot of Pilotview, showing assignments of individual locomotives to trains. Pilotview allows the user to click on locomotives, trains, and locomotive-to-train assignments to access additional information.

Figure 7 - Simulated train delay versus fleet size, compared to historical performance, demonstrating close agreement. After calibrating the model on one dataset, this plot was produced on a new dataset without any further calibration.
The model also has the ability to balance loads across shop locations when routing power to shop. Shop routing is a particularly sophisticated feature of the model. It uses adaptive learning to estimate the time required to get a locomotive to each shop (given all downstream events), and to estimate the backlog at each shop. We can then introduce a penalty to reduce these backlogs. Figure 9 shows the total backlog across all the shop locations as a function of the time within the simulation. The model can do little to reduce backlogs early in the simulation, which are largely a result of initial conditions, but with a higher penalty, the backlogs are reduced as the simulation progresses (note that this behavior is learned over the course of about 50 iterations).

These features make it possible to tune the model so that it can achieve realistic behaviors. For example, it is important that the model handle consist breakups and shop routing in a realistic way if it is going to be used for fleet sizing. Ignoring these important operational issues would allow the model to achieve levels of utilization higher than what could be possible in the field, a common problem with the use of classical optimization models. In addition, these features mean that it is possible to perform strategic planning studies that use highly realistic models of railroad operations. This realism was essential to capturing the productivity of locomotives. Without this, the model could not be used to model fleet size.

At this point, we concluded that the model was calibrated, and responded in a smooth and consistent way to changes in the input parameters.

**Strategic planning**

The most important strategic planning question at Norfolk Southern involved estimating the appropriate fleet size and mix given a projected train schedule. NS had used a simple regression model to estimate fleet size, but management came to feel that inefficiencies were baked into this model. The
The development of PLASMA was motivated by the desire to have an engineering solution that could adapt in a realistic way to assumptions about train schedules, fleet size and mix, and network performance.

When the model is used for strategic planning, all locomotives start in a “super source” node. We do not have to specify where locomotives are initially, and we do not even have to specify the fleet mix, although we are allowed to do so. The model then figures out where to first position each locomotive at the beginning of the planning horizon by using the value function approximations for the starting period. After this decision is made, the adaptive learning logic assigns power to trains over a planning period (typically a month). The use of an optimization-based modeling strategy means that the model simulates a well-trained group of locomotive planners.

Figure 10 illustrates how the model is used to estimate fleet size. The model is first used to create a train delay curve (total delay as a function of fleet size) for the current year. Then, a projected train schedule is created for some period in the future, after which the model is used to create a new delay curve. If we would like to maintain the same level of train delay, we can simply pick off the required number of locomotives. If we do not constrain the model to a fixed proportion of different locomotive types, the model will also specify the fleet mix.

The model can be used to perform different types of policy studies. Figure 11 illustrates an analysis of the effect of changes in average train speed. Train delay curves were generated for a base case, and then for six scenarios where the average train speed was varied. We note that the curves are quite consistent and well-behaved, simplifying the task of identifying the correct fleet size.

While Norfolk Southern has primarily used the model for fleet sizing studies, it can be used for other questions such as quantifying the effect of changes to the train schedule, changes in interchange points to foreign railroads, and changes in the size and location of maintenance shops.
Operational forecasting

An operational forecasting model produces a plan over perhaps a five to seven day horizon. The model is used to identify surpluses and deficits of power, and to anticipate locomotive repositioning and light engine moves (moving power without a train). Such a model requires that we know where the locomotives are initially. Thus, while the strategic planning model has to figure out where each locomotive should be at the beginning of the simulation, the operational forecasting model works from a live snapshot.

The operational forecasting model at NS runs in a production setting. After each forward sweep (over, say, a seven day horizon), the model would refresh the locomotive snapshot, as well as capture any changes to the train schedule. This process should repeat itself approximately once each minute (for a network comparable to that of Norfolk Southern). In the process, the model is constantly refining the value function approximations.

The operational forecasting model requires that the train schedule and locomotive snapshot be accurate (the strategic planning model does not require a locomotive snapshot). This is not a small request for a railroad. Norfolk Southern has been extensively testing and validating the operational forecasting model, but in the meantime the process has helped to identify areas where data reporting needs to be more accurate. This will be realized through upgrades to the information systems and improved reporting procedures. A byproduct of this implementation has sparked a major revision of their data collection and reporting process for locomotives. This is a familiar experience, where the process of
implementing advanced decision support systems has the effect of raising the bar on the quality of information systems. Usage at Norfolk Southern

Prior to the development of the LARS with the PLASMA planning engine, Norfolk Southern used a simple regression model to estimate future locomotive fleet-size requirements. This model looked like

\[
\text{Locomotive Fleet-Size} = \theta_0 + \theta_1 \text{(total carloads)} + \theta_2 \text{(system average train speed)}
\]

Marketing provided an aggregated forecast of the total carloads of freight that would move over the network in a future period. System train speed would be estimated based upon recent historical observations and general expectations of operating performance trends. This model had been tuned over the years, and while the limitations of the linear regression for forecasting were acknowledged, senior management had no better alternatives.

The importance of getting the right fleet size cannot be underestimated. Norfolk Southern currently runs a road fleet of over 2,000 locomotives with new ones costing well over $2 million apiece. If there are too few locomotives, customer service will deteriorate as train delays increase, leading to a loss of competitiveness in key markets. In addition to the large capital expenditure, excessive locomotives are expensive to maintain or store. Furthermore, excess locomotives soften the impact of poor execution which may hide issues in other areas. Too many locomotives may be difficult to diagnose as locomotive utilization may not drop significantly due to increased number of locomotives per train.

Management continued to press for better forecasting options. Intuitively, adding the first additional locomotive to the fleet will have a greater impact than, say, adding the 200th, but the linear model did
not differentiate. Furthermore, the linear regression model perpetuated historical locomotive performance even as operations improved. Train velocity was the only performance variable in the old model that could be adjusted for future periods. In contrast, as depicted in figure 7 (and figure 10), the PLASMA model built within LARS predicts declining improvements as the fleet size is increased, as we would expect.

The extensive efforts behind the multi-year calibration process of LARS helped to build confidence in the estimates of fleet-size produced by the model. Still, it took time for management to begin to trust the recommendations of the model. LARS went into production right before the 2008 recession began while the marketing forecasts reflected continued growth into the foreseeable future. Quickly, however, Norfolk Southern used LARS to estimate the number of excess locomotives in the system given the forecasted reduced demand. With the target in mind, owned locomotives were stored or leased locomotives returned on a weekly basis.

Later, in 2010 as the economy slowly rebounded, the model played a key part in the decisions to release units from storage and eventually to acquire new locomotives. An accurate future estimate was essential as the economic conditions changed. Underestimating the required fleet size would result in train delays that impact service while overestimating the fleet not only results in millions of dollars in asset costs, but also results additional shop costs and yard congestion due to storing unneeded locomotives.

One of the biggest advantages of the LARS model is that it considers the traffic mix. Due to routing and loading variations, not all carloads (or intermodal units) are equal. Norfolk Southern has developed a process utilizing several models to create a full train schedule based upon the carload forecast produced by marketing. This has been very important recently as coal volumes have decreased while intermodal volumes increased. Accurately assessing this impact has been essential in managing the fleet.

LARS produces not only an estimate of the number of locomotives required to maintain service, but also provides a range with a desired confidence level. This range was communicated to senior management to help them balance upside and downside risks. This analysis has been instrumental in determining the optimal size of a surge fleet – stored locomotives that can be pulled out and utilized with weeks of advanced notice.

Since 2008, Norfolk Southern has been increasingly relying on the estimates produced by LARS. LARS also makes it possible to understand how the fleet size is impacted by variations in train schedules, variations in power transfer time between trains, policies on foreign online power, variations in train velocity, and changes in locomotive technology such as increased horsepower.

**Benefits of LARS**

LARS provides a number of features that are not offered by the linear regression model previously used at Norfolk Southern (given above). These include:

- Estimates of fleet size and mix depend not just on aggregate tonnage forecasts, but the type of freight being moved and its spatial characteristics which are not captured by the linear model.
Fleet size estimates also depend on operating characteristics such as set-off times, train speed (which may change as a result of specific track improvements), and service levels.

- Estimates of fleet size depend on the operating plan, which includes the train schedule and blocking plan, which are completely ignored by the regression model.
- LARS can produce estimates of fleet size for a specific business unit such as coal, automotive and intermodal.
- Unexpected events can be simulated (e.g. a track shutdown due to snow or an accident), making it possible to estimate the fleet size needed to handle such events. LARS models intelligent repositioning in response to major changes such as this.
- Fleet size estimates reflect the flows of locomotives from neighboring railroads (“foreign power”).
- LARS can capture the effect of changes in the mix of different types of locomotives.
- Simulations using LARS capture the nonlinear effect of changes in fleet size. Additional locomotives have diminishing returns, while the reverse is true as the fleet is reduced.
- LARS can be recalibrated each year, for different periods of the year. By contrast, a regression model must be estimated based on many years of data.

These benefits come, of course, at a cost. LARS requires substantial efforts to prepare the data and calibrate the model. For example, it is not enough just to estimate the total tonnage for a future year, but this forecast has to be broken down by origin, destination and commodity type, and then translated into train tonnages. The value of a detailed engineering model such as PLASMA is that it is sensitive to a wide range of inputs, but this also means that errors in these inputs can translate to errors in model outputs. The finding at Norfolk Southern is that errors in the tonnages between different cities largely cancel out. Of course, if we systematically overestimate the tonnage by 5 percent over the entire system, then the model will need more locomotives to move this traffic, but that would be true with an aggregate model. However, the time required to prepare these detailed inputs should not be underestimated.

A separate issue is that the high level of detail comes with a price in terms of computation time. A single run of the model requires several hours of CPU time, and graphs such as those depicted in Figure 7 can require several days to generate. However, the impact on operating costs, performance and network robustness are dramatically larger than the cost of developing and maintaining the system.

**Conclusions**

Approximate dynamic programming offers a novel modeling and algorithmic strategy that combines the flexibility and realism of simulation with the intelligence of optimization. Classical optimization models have offered the promise of better decisions, but the technology has required the use of major simplifying assumptions. As a result, the “savings” produced by such models are often a by-product of simplified models rather than intelligent decisions.

PLASMA has been shown to produce high quality, accurate solutions to strategic and tactical planning problems at Norfolk Southern. Furthermore, it has shown very promising results for operational forecasting. It is the first optimization-based model of locomotives for a North American freight railroad that calibrates accurately against history, making it useful as a tool for fleet sizing, one of the most demanding strategic planning problems. The technology allows locomotives and trains to be modeled at
an extremely high level of detail. Train delays can be modeled down to the minute. The model can simultaneously handle consist breakups and shop routing, while also planning the empty repositioning of power. In addition, it can handle uncertainties in transit times, yard delays and equipment failures in a simple and intuitive way. The entire methodology is based on first principles, and as a result avoids the need for heuristic rules that have to be retuned as the data changes.

Acknowledgements: We would like to thank the following people who have made significant contributions to this project: Don Grabb, Ed Courney, Junxia Chang, Brian Wilker, and Jermaine Wilkinson. The early stages of this project were managed by Ajith Wijeratne (who was the first to introduce the train delay curves), and the first author would like to thank the very early support of Roger Baugher who recognized the potential of approximate dynamic programming for rail operations. The research behind this work has been supported over the years by the Air Force Office of Scientific Research and the National Science Foundation.

References


