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

Warren B. Powell, Belgacem Bouziaine-Ayari
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544
(powell@princeton.edu, belgacem@princeton.edu)

Coleman Lawrence, Clark Cheng, Sourav Das, Ricardo Fiorillo
Norfolk Southern Corporation, Atlanta, Georgia 30309
(coleman.lawrence@nscorp.com, clark.cheng@nscorp.com, sourav.das@nscorp.com, ricardo.fiorillo@nscorp.com)

For decades, locomotive planning has been approached using the classical tools of mathematical programming; the result has been very large-scale integer programming models that are beyond the capabilities of modern solvers but still require a host of simplifying assumptions that limit their use for analyzing important planning problems. The primary interest of Norfolk Southern was in developing a model that could assist it with fleet sizing. However, the cumulative effect of the simplifications required to produce a practical integer programming formulation resulted in models that underestimated the required fleet. We use the modeling and algorithmic framework of approximate dynamic programming, which uses an intuitive balance of simulation and optimization with feedback learning, to produce a highly detailed model that calibrates accurately against historical metrics. The result was a model that can be used to plan fleet size and mix, be sensitive to a wide range of operating parameters, and adapt to many scenarios.

Keywords: transportation: rail; programming: dynamic; approximate dynamic programming; locomotives.

History: This paper was refereed. Published online in Articles in Advance.
the design of specialized algorithms to solve integer programming formulations for locomotive scheduling. Ziarati et al. (1997, 1999) present a new branch-and-cut algorithm. Ahuja et al. (2005) describe a heuristic based on very large-scale neighborhoods to find high-quality schedules for locomotives; this heuristic considers the grouping of locomotives into \textit{consists} required to pull a train and the desire for weekly patterns in the flows of locomotives. Vaidyanathan et al. (2008) provide a detailed model of the locomotive routing problem, capturing a number of operational constraints with an adaptation of their large neighborhood search strategy; see Vaidyanathan and Ahuja (2008) for additional experimental work. Cordeau et al. (2000, 2001) describe the use of Benders decomposition for the simultaneous assignment of cars and locomotives. However, we must also recognize major advances to general-purpose integer programming solvers, such as CPLEX and Gurobi, that have occurred since 2005. Later, we report on experiments with CPLEX 12 running on a single machine with 1 TB 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 an information system developed at Norfolk Southern called the Locomotive Assignment and Routing System (LARS). As of this writing, LARS 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. PLASMA has been undergoing extensive user-acceptance testing while Norfolk Southern has been upgrading its information systems to improve the accuracy of some of its data.

**Locomotive Operations**

A locomotive can be described by a host of attributes, including horsepower and tractive effort (tractive effort captures the ability of steel wheels to grab the rail when the train is starting up), its owning railroad, its maintenance status (e.g., number of days until its next federally mandated maintenance appointment), and equipment details such as communications gear (for coordinating multiple locomotives). Locomotives typically are bundled together into consists of two to four locomotives that are needed to pull a particular train. Because forming a locomotive consist requires connecting and testing the cables that allow the set of locomotives to work as a single unit, it is a time-consuming process. Trains often arrive from a neighboring railroad using locomotives owned by that railroad (i.e., foreign power). Foreign locomotives are usually returned to the owning railroad; however, railroads other than the owning railroad can sometimes use these locomotives within negotiated limits. If the locomotive is returned, the return must be handled through predefined exchange points.

The trains also have numerous attributes. The number of locomotives needed to pull a train depends on the weight of the train, its speed requirements (e.g., merchandise trains need to move more quickly than coal and grain trains), and the steepest grade that the train must navigate. Trains also have different service priorities, and a scheduler has to consider whether enough locomotives are available to move all the trains on time to ensure that service priorities are met.

Shop routing (i.e., moving a locomotive to a maintenance shop) is one of the most complex issues facing a locomotive manager. Locomotives require government-mandated inspections on a periodic schedule (typically every 92 days). If a locomotive is not inspected on time, it must be turned off and towed to the maintenance shop. For some trains, simply adding an extra locomotive and using it to move the locomotive requiring maintenance to its shop appointment is possible; however, this is an unproductive use of the extra locomotive. The scheduler has to decide whether to get the locomotive to the maintenance shop early, which has a productivity cost, or risk that it may arrive at the shop late (e.g., because it has missed a critical connection). In addition, multiple shop locations may be available to service the locomotive. A planner has to anticipate the number of locomotives that are scheduled at each shop to balance the loads (at the shops) when making an assignment of a locomotive to a train.

A separate issue that is widely discussed, but rarely resolved, concerns the various types of uncertainty that plague locomotive operations. These include

- \textit{Transit-time delays}: These delays can range from as long as 6 to 12 hours for the shorter movements of
an Eastern railroad to more than a day for the long movements of the Western railroads. Random transit times are without question the most significant source of uncertainty in rail operations.

- **Dynamic schedule changes**: Schedulers must 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 provide projections of when a locomotive will be ready to leave a shop; however, these are just estimates and schedulers have to call the maintenance facility to determine if a locomotive will be ready for a particular train.
- **Equipment failures**: Locomotives may fail unexpectedly, resulting in additional uncertainty.

We designed the model we present in this paper to handle uncertainty; however, production applications of the model have yet to exploit this capability. The difference between running the model assuming perfect information versus modeling uncertainty is actually quite minor.

### 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 (see 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 between terminals. The figure depicts three types of locomotives in different shades of gray. This classical model must be modified to handle issues such as the coupling and (or) uncoupling of locomotives to move a train and routing to shop. However, perhaps one of the most difficult challenges is modeling 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 (e.g., at 2 PM, 4 PM, and 6 PM), when the train may have to be delayed for a time (e.g., 37 minutes) to handle a late-arriving inbound train. We found that rounding departure times to the nearest two hours, for example, produced scheduling errors that were unacceptable to the railroad.

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 lengthens, we have to model the cascading of train delays. With a horizon of as few as four days, the run times already exceed 50 hours.

### Model Based on Approximate Dynamic Programming

Approximate dynamic programming (ADP) is a modeling and algorithmic strategy that decomposes decisions...
Locomotives
Horsepower
4,400
4,400
6,000
4,400
5,700
4,600
6,200

Locomotive
value
functions

Train “reward” function
Trains Downstream
yard
Aggregate
value
functions

ADP captures the value of locomotives in the future using a value-function approximation. We approximate the value function with piecewise linear, separable functions (see Figure 3). This allows us to solve the assignment problem (see Equation (1) in the appendix) as an integer program, which can be solved by commercial solvers in a few seconds, even when optimizing more than 2,000 locomotives for the entire railroad. We can use this integer program to estimate the marginal value of each locomotive and use these marginal values to estimate the piecewise linear functions. The appendix includes a brief mathematical summary of the problem; Bouzaïne-Ayari et al. (2014) give a complete description of the mathematical model and algorithm.

The assignment model (see Figure 3), 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 a train (which may be delayed); we can model consist breakup and formation, routing of foreign locomotives, routing of locomotives toward shop locations, and the need for specific equipment types within a consist. For example, we can model that two locomotives are available at 12:42 PM, whereas a third locomotive will not arrive until 1:37 PM, for a train that should have departed at 12:50 PM but which will have to be delayed until the last locomotive has arrived.

A train reward function captures the value of assigning power to a train. The highest marginal value is given to the power needed to move the train, with a somewhat lower marginal value given for power that brings the train from the minimum horsepower needed to move the train, up to the goal horsepower, which allows the train to move fast enough to meet service goals. There is a slight negative marginal value given to excess power, which discourages using more power than needed but allows it to meet downstream power requirements.

The value-function approximations are a distinguishing feature of ADP. Rather than optimizing over a long horizon all at the same time, they approximate the downstream impact of a decision now, reducing a single large problem into a series of much smaller problems (much the way people break down large problems). The model internally learns them by repeatedly simulating the process of assigning locomotives to trains (see Figure 4).

In Figure 4, the gray boxes and 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 four-hour time steps, although the timing of all assignments of locomotives to trains are calculated 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 one month for strategic planning problems; however, 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 and at each location. These marginal estimates are then smoothed into piecewise linear value functions at two levels of aggregation for each yard. Over repeated learning iterations, 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, because the value functions at the beginning of the horizon capture the marginal value of each locomotive type over the entire horizon.

The process of stepping forward in time and solving sequences of relatively small problems is the reason that ADP is often referred to as an optimizing simulator. Although the integer program in Figure 3 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 in which a decision anticipates the future. For example, although a train may require only two locomotives, 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. Figure 5 describes the steps of the ADP algorithm.

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 (Bouzaïne-Ayari et al. 2014) has shown that this produces very robust dispatch policies. However, Norfolk Southern has not as yet adopted policies that were trained using such a sampling policy.

Model Calibration and Validation

We carefully calibrated the model against several years of historical performance. This required us to painstakingly examine detailed assignments and compare high-level performance metrics. The process required that we iteratively identify and correct data errors, enhance the model, and make occasional improvements to the basic algorithm. For example, we identified the need to use two layers of aggregation in the value-function approximations through this process (see 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 handling foreign powers and the rules for consist formation.

In addition to carefully examining 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 almost perfectly match locomotive productivity. For example,
accurately capturing the costs and time required to break up locomotive consists is important. If we ignore this component, we would overestimate 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 produced delays that were an order of magnitude larger than delays in the historical data, largely because of data errors that caused locomotives to be hopelessly out of position. Matching historical performance by simply tuning parameters within the model was not possible. Experienced schedulers had to examine the detailed assignments. We were able to simplify this process by using a powerful diagnostic tool called Pilotview (see Figure 6), which 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 revising the model so that it closely matched historical performance, we repeated the exercise with a new data set. The result is the curve shown in Figure 7, which very closely matches the historical delay curve 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 difficult, because it requires that we have the ability to continuously

Figure 6: A snapshot of Pilotview shows 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: The graph shows simulated train delay versus fleet size, compared to historical performance, and demonstrates their close agreement. After calibrating the model on one data set, we produced this plot on a new data set without requiring additional calibration.
model train delays. 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.

Other forms of model validation involved testing the model’s sensitivity 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’s objective function, which captures repositioning costs and penalties for delayed trains.

The model can also balance loads across shop locations when routing power to a maintenance shop. Shop routing is a particularly sophisticated feature of the model. It uses adaptive learning to estimate the time required to move a locomotive to each shop (given all downstream events) and to estimate the backlog at each shop. We can then introduce a congestion 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. Early in the simulation, the model can do little to reduce backlogs, which are largely a result of initial conditions; however, when we use a higher congestion penalty, the backlogs decrease as the simulation progresses (note the model learned this behavior 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, if the model is to be used for fleet sizing, it must handle consist breakups and shop routing in a realistic way. Ignoring these important operational issues would allow the model to achieve levels of utilization higher than the levels possible in the field, a common problem when using classical optimization models (these simplifications result in phantom savings, where the model claims the railroad can operate with fewer locomotives than would be possible). In addition, these features mean that it is possible to perform strategic planning studies that use highly realistic models of railroad operations. This realism is essential to capturing the productivity of locomotives. Otherwise, the model could not be used to model fleet size. At this point in our work, we concluded that the model was calibrated and was responding 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. Norfolk Southern had used a simple regression model to estimate fleet size, but management realized that inefficiencies were inherent in this model. The development of PLASMA was motivated by management’s
desire for an engineering solution that could realistically adapt 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 (a “node in the sky”); that is, we do not have to specify where locomotives are initially and we do not have to specify the fleet mix, although we are allowed to do so. The model then determines where to initially position each locomotive at the beginning of the planning horizon by using the value-function approximations for the starting period. After the model has made this decision, 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. It is used initially to create a train delay curve (i.e., 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, it 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. We generated train delay curves for a base case and then for six scenarios in which we varied the average train speed. We note that the curves are consistent and well behaved, simplifying the task of identifying the correct fleet size. Although 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 a horizon of perhaps five to seven days. The model is used to identify surpluses and deficits of power, and to anticipate locomotive repositioning and light engine moves (i.e., moving power without a train). Such a model requires that we know where the locomotives are initially. Thus, although the strategic planning model must determine where each locomotive should be at the beginning of the simulation, the operational forecasting model works from a live snapshot, which gives the status of each locomotive in the system.

The operational forecasting model at Norfolk Southern runs in a production setting. After each forward sweep (e.g., over a seven-day horizon), the model refreshes the locomotive snapshot and captures any changes to the train schedule. This process should
repeat itself approximately once each minute in a network comparable to that of Norfolk Southern. During this process, the model constantly refines 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 a major task for a railroad. Norfolk Southern has been extensively testing and validating the operational forecasting model; the process has helped it to identify areas in which its data reporting needs to be more accurate. The railroad will realize these accuracy improvements by upgrading its information systems and improving its reporting procedures. A by-product of this implementation has sparked a major revision of Norfolk Southern’s data collection and reporting process for locomotives. This experience, in which the process of implementing advanced decision support systems has the effect of raising the bar on the quality of information systems, is typical.

Usage at Norfolk Southern

Prior to the development of LARS with the PLASMA planning engine, Norfolk Southern used a simple regression model to estimate future locomotive fleet-size requirements, as the following equation shows:

\[
\text{Locomotive fleet size} = \Theta_0 + \Theta_1(\text{total carloads}) + \Theta_2(\text{system average train speed})
\]

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

We cannot underestimate the importance of determining the correct fleet size. Norfolk Southern currently runs a road fleet of over 2,000 locomotives and a new locomotive costs well over $2 million. An insufficient number of locomotives results in deteriorating customer service as train delays increase and leads to a loss of competitiveness in key markets. In contrast, an excessive number of locomotives is a large capital expenditure and is expensive to maintain or store. Furthermore, having too many locomotives can mitigate the impact of poor execution, which may hide issues in other areas. This may be difficult to diagnose because locomotive utilization may not drop significantly as a result of an 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, for example, adding the 200th; however, the linear model could not consider this factor. 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 Figure 7 (and Figure 10) depict, the PLASMA model built within LARS predicts declining improvements as the fleet size increases, as we would expect.

The extensive efforts behind the multiyear calibration process of LARS helped to build confidence in the fleet-size estimates produced by the model. However, management needed time to begin to trust the model’s recommendations. LARS went into production immediately before the 2008 recession began, a period in which the marketing forecasts reflected continued growth into the foreseeable future. Norfolk Southern was able to quickly use LARS to estimate the number of excess locomotives in the system, given the forecasted reduced demand. These estimates were then used to adjust the fleet each week. In addition, the model was used to estimate fleet size and mix over extended horizons given forecasted schedules.

Later, as the economy slowly rebounded in 2010, the model played a key part in the decisions to release locomotives from storage and to eventually acquire new locomotives. An accurate future estimate was essential as economic conditions changed. Underestimating the required fleet size would result in train delays that impact service; overestimating the fleet would result in millions of dollars in asset costs and additional shop costs, as well as yard congestion because of storing unneeded locomotives.

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

LARS produces an estimate of the number of locomotives required to maintain service, and it also provides a range with a desired confidence level. We communicated this range to senior management to help it balance the 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 used with a few weeks of advance notice.

Since 2008, Norfolk Southern has increasingly relied on the estimates produced by LARS. LARS enables management and staff to understand how the fleet size is impacted by variations in train schedules, variations in power transfer time between trains, policies on the management of locomotives belonging to other railroads, variations in train velocity, and changes in locomotive technology (e.g., increased horsepower).

**Benefits of LARS**

LARS provides a number of features that the linear regression model Norfolk Southern used previously does not offer. These include the following.

- Estimates of fleet size and mix depend not just on aggregate tonnage forecasts but on the type of freight being moved and its spatial characteristics, which the linear model cannot capture. Fleet-size estimates also depend on operating characteristics, such as the time to remove locomotives from a train (known 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 the regression model ignores.

- LARS can produce estimates of fleet size for a specific business unit, such as coal, automotive, or intermodal.

- Unexpected events can be simulated (e.g., a track shutdown because of 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 and unexpected events.

- Fleet-size estimates reflect the flows of locomotives from neighboring railroads (i.e., foreign powers).

- 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, whereas the reverse is true as the fleet is reduced.

- Each year, LARS can be recalibrated for different periods of the year. By contrast, a regression model must be estimated based on many years of data.

These benefits have a cost. LARS requires substantial effort to prepare the data and calibrate the model. For example, estimating the total tonnage for a future year is not sufficient; this forecast must 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; however, errors in these inputs can translate to errors in model output. Norfolk Southern found that errors in the tonnages between different cities largely cancel out. Of course, if we systematically overestimate the tonnage by five percent over the entire system, then the model will need more locomotives to move this traffic, which would also 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 computation-time cost. A single run of the model requires several hours of CPU time, and graphs such as those depicted in Figure 7 can take several days to generate. However, the impact on operating costs, performance, and network robustness is dramatically large relative to the cost of developing and maintaining the system.

**Conclusions**

ADP offers a novel modeling and algorithmic strategy that combines the flexibility and realism of simulation with the intelligence of optimization. Classical optimization models offer the promise of better decisions; however, the technology requires 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.
We have shown that PLASMA can produce high-quality, accurate solutions to strategic and tactical planning problems at Norfolk Southern. Furthermore, its results are very promising for operational forecasting. To our knowledge, it is the first optimization-based model of locomotives for a North American freight railroad that calibrates accurately against historical data, 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 repositioning of locomotives to other terminals. In addition, it can handle uncertainties in transit times, yard delays, and equipment failures in a simple and intuitive way. The methodology is based on a formal mathematical model, which guided the design of rigorous algorithms; as a result, it avoids the need for heuristic rules that have to be retuned as the data change.

Appendix

Mathematically, we can write this problem in the form of a dynamic program. Let $R_t = (R_{td})_{a \in A}$, where $R_{ta}$ is the number of locomotives with attribute $a$ at time $t$, where $a$ captures the type of locomotive, its location, maintenance status (e.g., the next time it is due for shop maintenance), special equipment (e.g., flush toilets and communication), and whether it is attached to other locomotives in a consist. Let $D_t$ be the set of trains waiting to be moved (i.e., 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 $S_t$ at time $t$, then we can use Bellman’s optimality equation to write

$$V_t(S_t) = \min_{x \in X_t} \left( \sum_{a \in A} \sum_{d \in D_t} c_{ad} x_{ad} + V_{t+1}(S_{t+1}(S_t, x_t)) \right), \quad (1)$$

where $x_{ad}$ 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 (cost must consider the need to break apart consists and the appropriateness of a particular type of locomotive for a particular type of train). The state of the system is $S_{t+1}(S_t, x_t)$, which results from the decision vector $x_t$.

Bouzaïne-Ayari et al. (2014) provide a detailed mathematical formulation of the ADP model and algorithm. At the heart of the model is the locomotive assignment subproblem, where we assign locomotives to trains at a particular point in time (see Figure 3); this subproblem 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

$$X^*_t(S_t) = \arg \min_{x \in X_t^*} \left( \sum_{a \in A} \sum_{d \in D_t} c_{ad} x_{ad} + V_{t+1}(S_{t+1}(S'_t, x_t)) \right). \quad (2)$$

This is the problem we solve at iteration $n$, time $t$, using the value-function approximation $V_{t+1}(S_{t+1})$ from the previous iteration. If $x^*_t$ is the optimal solution, we obtain the next state using $S^*_{t+1} = S_{t+1}(S'_t, x^*_t)$, which we obtain by simulating from state $S^*_t$ at time $t$ to state $S^*_t$ at time $t+1$, during iteration $n$.

Acknowledgments

We 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 thanks the very early support of Roger Baugher, who recognized the potential of ADP for rail operations. The research behind the approximate dynamic programming algorithms was supported by the Air Force Office of Scientific Research [Grants AFOSR-FA9550-05-1-0121 and AFOSR-FA9550-06-1-0496] and by earlier grants from the National Science Foundation.

References


Verification Letter
Jerry W. Hall, Vice President Network and Service Management, Norfolk Southern, 1200 Peachtree St. NE, Atlanta, GA 30309, writes:

“Norfolk Southern has implemented the Princeton Locomotive And Shop Management system (PLASMA) as part of an information system called the Locomotive Assignment and Routing System (LARS).

LARS went into production in 2008. Since then, it has been used for strategic locomotive fleet planning. Specifically, LARS has been used to estimate the number of excess locomotive during economic downturns while determining the number of locomotives to be acquired to maintain service and meet future demand during economic growth. LARS has been instrumental in determining the optimal size of a surge fleet—stored locomotives that can be pulled out and used within a few weeks of advance notice.

LARS has a powerful what-if capability which enables management and staff to understand how the fleet size is impacted by variations in train schedules, power transfer time between trains, policies on the management of locomotives belonging to other railroads, variations in train velocity, and changes in locomotive technology.

Norfolk Southern has heavily relied on the estimates produced by LARS. Therefore, LARS has become an integral part of the company’s network and asset planning processes and it provides Norfolk Southern with a competitive advantage in the ever-changing marketplace.”

Warren B. Powell is a professor in the Department of Operations Research and Financial Engineering at Princeton University, and he is the director of CASTLE Labs, which develops advanced models and algorithms for applications in transportation and logistics, energy, health, and finance. Professor Powell is the author of Approximate Dynamic Programming: Solving the Curses of Dimensionality, which was motivated by this research. He has coauthored over 200 publications and is a fellow of INFORMS.

Belgacem Bouzaïene-Ayari is a senior research staff in CASTLE Labs at Princeton University, where he has worked since 1996. Belgacem is an expert in the development of large-scale optimization models and was the lead developer for PLASMA, serving as the critical interface between academic research and Norfolk Southern.

Coleman Lawrence is currently in the 19th year of his career at Norfolk Southern. For the last seven years, he has held the position of senior director operations and locomotive control in Atlanta. Prior to this position he held a variety of positions in Norfolk Southern’s operations, finance, and marketing divisions. In his capacity as the primary manager of Norfolk Southern’s locomotive fleet, he has acted as the primary business partner in the development of the company’s LARS system based on the PLASMA model. He earned BS and MS degrees in civil engineering at Virginia Military Institute and the Massachusetts Institute of Technology, respectively, and an MBA from the University of Virginia.

Clark Cheng is the director of the Operations Research Department at Norfolk Southern, which develops optimization, simulation, and forecasting models and decision support systems for operations and asset planning and management. During his two-decade railroad career, he has participated in developing and implementing models, algorithms and decision support systems for network planning and optimization, car scheduling, locomotive and railcar fleet planning, demand forecasting, rail yard simulation and optimization, crew planning, line capacity, blocking optimization, train schedule optimization, locomotive dispatching, and empty railcar distribution. His department also develops advanced analytics systems to improve business performance and reduce operating costs.

Sourav Das worked as a senior data analyst in the Operations Research Department at Norfolk Southern. Dr. Das has worked on developing and implementing models for locomotive fleet planning, car scheduling, demand forecasting, rail yard simulation and optimization, locomotive dispatching, and empty railcar distribution. Currently, he is working as an independent consultant.

Ricardo Fiorillo is the manager of operations research at Norfolk Southern. He has worked in the locomotive asset management arena for over 15 years. Ricardo currently manages Norfolk Southern’s efforts to apply the PLASMA model to locomotive fleet-size planning, short-term tactical forecasting, shop routing, and real-time locomotive-to-train assignments.