Big Data and Railroad Analytics

David Hunt and Jason Kuehn, ***Oliver Wyman***; **David.Hunt@oliverwyman.com and Jason.Kuehn@oliverwyman.com**

The new movie "Moneyball," starring Brad Pitt, tells the story of how a low-budget baseball team was able to effectively compete with the wealthiest teams through better analytics. By focusing on metrics that lead to victories, the Oakland Athletics avoided high-priced players and signed undervalued players whose statistics translated into more wins.

Railroads may well rival baseball in terms of the amount of data collected and the number of statistics published. The earliest railroads collected financial and operating data to facilitate improvements. As easily as you can learn that outfielder Hugh Duffy had a .440 batting average for the Boston Beaneaters in 1894, you can also discover that there were 9,021 miles of track in the US in 1850, or that the Pacific Railroad carried an estimated value of $50 million in freight over the Great Plains in 1880.

With the widespread use of computers in the latter part of the 20th century, railroad analytics rapidly expanded from producing tables of facts to performing in-depth analysis that helped drive decisions. It also led to exponential growth in the amount of data that could be collected and stored. It was now possible to analyze millions of waybill records to build a better operating plan, or plow through mountains of financial data to improve the bottom line. National data sources included reliable time-series economic data from the Surface Transportation Board/ Interstate Commerce Commission, accident and incident data from the Federal Railroad Administration, and numerous statistics from the Association of American Railroads, including the "Green Book" which became the bible of railroad statistics.

Despite this long history of railroad analytics, we are really just on the precipice of what can be done. Today, railroads collect enormous quantities of data through GPS, AEI readers, electronic data interchange, video inspections, handheld field tablets, and many other sources. The data are growing not only in quantity, but also in quality as they are more frequent and more precise. The problem is that the tools to extract information from these data are not keeping pace with the data explosion. We are quickly exceeding the limits of spreadsheet analysis.

An example of a tool specifically designed to handle massive railroad databases is the Traffic Flow Analyzer (TFA); built for CSX by Oliver Wyman. The TFA archives three years of car and train movement records in a custom designed data warehouse. The TFA provides data extraction and mapping tools that can be used to generate traffic density maps, geographic pie maps of yard activity, and detailed reports. What is really unique about the TFA is that it does not model rail routes, but it stores the actual route taken by each railcar and each train on every trip. This provides tremendous value for rigorous engineering studies of loads on bridges and line segments, precise financial calculations involving ton-mile and car-mile statistics, federal compliance when reporting exact hazardous material routings, and many other uses throughout the company.

Analyzing big data is not just about databases growing bigger, it can also be about merging databases to extract even more useful intelligence. Imagine detailed network mappings of track layouts merged with inspection data on rail, tie, and ballast conditions also merged with traffic density and tonnage data. This data cocktail would be very powerful in defining capital budgets and prioritizing work programs. Another example would be the merging of financial and operational data to allow managing revenues (as opposed to trains) through network bottlenecks. By managing revenues, it becomes easier to identify the hurdle points that trigger capital investments for capacity expansion, or to build an operating plan that considers revenue maximization.

Tomorrow, the quantity and the quality of railroad data will be even greater than today, and the demand for timely information from these data will also increase. Just like in "Moneyball," the railroads that are best able to use analytics to improve decision making will be the ones that end up victorious. It is up to the OR community to deliver the right tool sets to make this possible.