Analytical Approaches to Validating Railroad Fuel Saving Technology

By:
Wayne A. Kennedy, Kennedy Consulting
Robert Stevens, First Analytics
Bret Lanz, Kansas State University
Spencer Maynes, Deflect Inc

Abstract

Class I railroads in North America have partnered with the Science Based Targets initiative during the last five years for very aggressive GHG emission goals by 2030 to 2034. Two of the six Class I's have committed to a net-zero target emissions profile by the year 2050. Historically, the rail industry has improved their fuel efficiency by an average of 1% per year going back two decades. If they continue on that glide slope to 2030, they will collectively have a 26% gap between where their current improvement trend will take them and where they will need to be in order to meet their SBTi goals. Clearly new approaches will be required in the immediate future to help close that gap.

In 2023, Class I railroads consumed approximately 3.6 billion gallons of diesel fuel including a small percentage of biodiesel, roughly 3% of the total. Total spending on this fuel was ~$11.9 billion, therefore a technology that can reduce this spend by 1% would save ~$119 million annually.

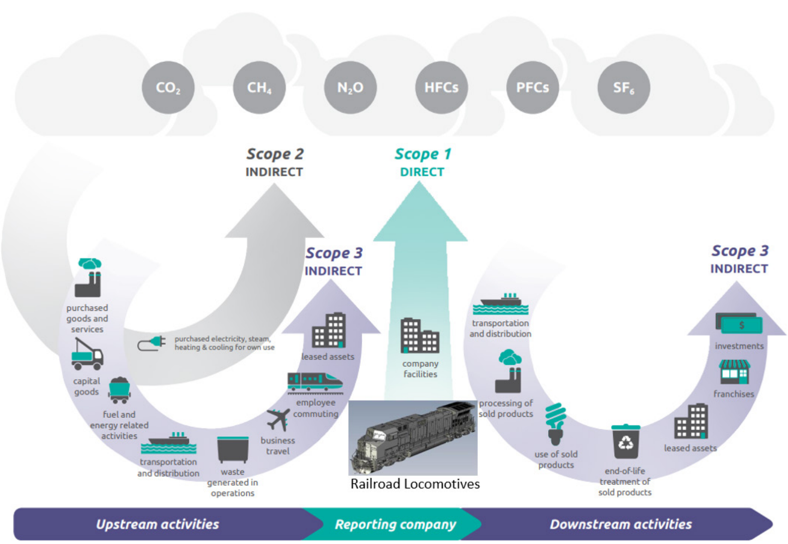
This paper focuses on various statistical analysis tools to validate the level of locomotive or train fuel savings for various technologies in the 1% to 3% range. The ability to prove fuel savings in a railroad operating environment at this "low" level of fuel savings is challenging for a variety of reasons. Inaccurate locomotive on-board fuel gauges is certainly one issue, another being the amount of inherent variability in overall fuel consumption in revenue freight railroading. There are so many different aspects which affect the amount of fuel for any given trip segment, all of which introduce a large amount of variability which makes proving low level fuel savings problematic at best, almost impossible at the worst. However, this paper outlines how it can be done, with appropriate design of experiments and statistical analysis.

SBTi Scope Emissions and Reduction Goals for the Industry

The Science Based Targets initiative (website sciencebasedtargets.org) was established in 2015 to help companies set emission reduction targets in line with climate science and the Paris Agreement goals. The Paris Agreement's long-term temperature goal is to keep the rise in mean global temperatures to well below 2 degrees Celsius above pre-industrial levels, and preferably limit the increase to 1.5 degrees Celsius. As of April 2024, there are almost 8,000 companies worldwide taking action.

Figure 1 shows the differing scopes and a brief description of each scope. Scope 1 covers direct emissions from owned or controlled sources, diesel fuel burned in locomotives in this case for rail. Scope 2 covers indirect emissions from the generation of purchased electricity, steam, heating and cooling consumed by the rail reporting company. Scope 3 includes all other indirect emissions that occur in a company's value chain.

There are two main metrics used for GHG emission measures; an absolute metric which is directly tied to how much diesel fuel is burned and is referenced in metric tons of CO_2 equivalent, and the more common measure based on intensity using either gross ton miles or revenue ton miles, as a base measure producing an intensity rate which is volume neutral. Example units are $MtCO_2e/GTM$ (gross ton mile).



Source: https://www.epa.gov/climateleadership/scope-3-inventory-guidance

Figure 1 – Emissions Scope Definitions

Table 1 is directly from the SBTi website and shows all Class 1 goals. The goal dates vary but are mostly centered around the 2030 timeframe. They have all chosen a "well below 2 degrees Celsius" goal for limiting long term global warming. Further, the CN, CPKC and UP have committed to a Net-Zero goal by 2050.

Both BNSF and UP railroads have chosen the Absolute goal measure, where future growth and volume increases will make their GHG reduction goals more challenging to achieve. Though, to-date, traffic volume loss of 14% as measured by Gross Ton Miles or GTMs from their SBTi base year of 2018 has helped both railroads make significant progress towards achieving their SBTi goals.

Class I Railroads and Description of GHG Emissions Reduction Goals

All railroads will need to submit more aggressive goals no later than 2025 to meet SBTi's preferred ambition moving from the current "well below 2 degrees Celsius" to a new "1.5 degree Celsius". The new strategy has been rolled out in response to increasing urgency for climate action and the success of science-based targets to-date.[1] Union Pacific was the first railroad to have their 1.5-degree Celsius goal approved by SBTi the end of March 2024, moving from a 26% to a 50% absolute reduction in GHG emissions.

Company Name	Full target language	Company Temperature	Scope	Target Value	Type	Base Year	Target	Date Published
BNSF Railway	BNSF Railway commits to reduce absolute scope 1 and 2 and well-to-wheel locomotive GHG emissions 30% by 2030 from a 2018 base year*. *The target boundary includes biogenic emissions and removals from bioenergy feedstocks.	Well-below 2°C	1+2	30%	Absolute	2018	2030	2023-05-25
Canadian National Railway Co	CN commits to reduce scope 1 and 2 GHG emissions 43% per gross ton miles by 2030 from a 2019 base year*. CN commits to reduce scope 3 GHG emissions from fuel and energy related activities 40% per gross ton miles by 2030 from a 2019 base year. *The target boundary includes biogenic emissions and removals from bioenergy feedstocks.	Well-below 2°C	1+2	43%	Intensity	2019	2030	2021-07-21
CSX Corporation	CSX commits to reduce scope 1 and 2 GHG emissions intensity 37% per million gross ton miles by 2029 from a 2014 base year.	Well-below 2°C	1+2	37%	Intensity	2014	2029	2020-01-01
Norfolk Southern Corporation	Norfolk Southern commits to reduce scope 1 and 2 GHG emissions 42% per million gross ton-miles (MGTM) by 2034 from a 2019 base year*. *The target boundary includes biogenic emissions and removals from bioenergy feedstocks.	Well-below 2°C	1+2	42%	Intensity	2019	2034	2021-07-29
Union Pacific Corporation	Union Pacific commits to reduce absolute scope 1 and 2 GHG emissions 50.4% by 2030 from a 2018 base year.* Union Pacific also commits to reduce scope 3 GHG emissions from purchased goods and services, capital goods, and fuel and energy-related activities 50.4% within the same timeframe. *The target boundary includes land-related emissions and removals from bioenergy feedstocks.	1.5°C	1+2	50%	Absolute	2018	2030	2024-03-28
CPKC	CPKC commits to reduce scope 1, 2, and 3 well-to-wheel locomotive GHG emissions 36.9% per gross ton-miles by 2030 from a 2020 base year.	Well-below 2°C	1+2+3	37%	Intensity	2020	2030	2023-11-23

Table 1 – SBTi Goals by Railroad

Figure 2 shows an equal weighting of the four Class I railroads that have chosen an emissions intensity goal. Note that improvement has stalled and actually worsened in the last two years and on average, they are now 15% above their SBTi glideslope goal. Current year is an estimate based on YTD fuel efficiency performance reported during quarterly earnings.

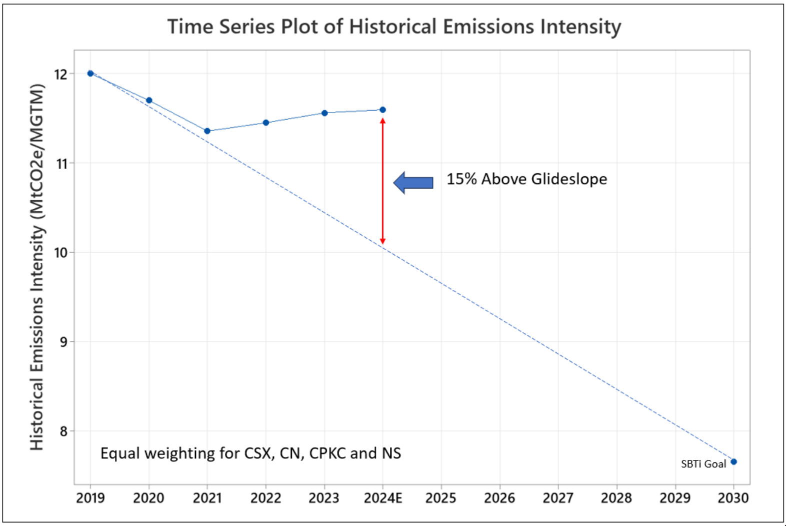


Figure 2 – Average Performance of Four Intensity Based Railroads Showing Glideslope to SBTi Goal

Figure 3 shows an equal weighting of the two Class I railroads that have chosen an absolute emissions goal. Given the significant drop in traffic volumes since the SBTi baseline year of 2018, it is not surprising that they are roughly on track for their required SBTi glideslope goal. Current year is an estimate based on YTD fuel efficiency performance reported during quarterly earnings.

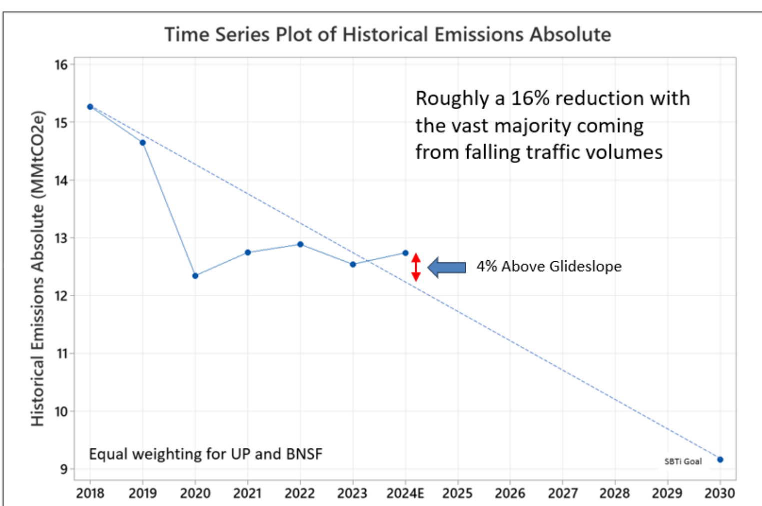


Figure 3 – Average Performance of Two Absolute Based Railroads Showing Glideslope to SBTi Goal

Historical Fuel Efficiency Industry Performance

The rail industry commonly uses a fuel efficiency measure of Gallons per Thousand Gross-ton-Miles which has been tracked for many decades. Figure 4 shows each of the seven Class 1 railroads going back to 2000. Notice the Canadian railroads have generally been the best performers and starting in 2013 both distanced themselves from their US peers.

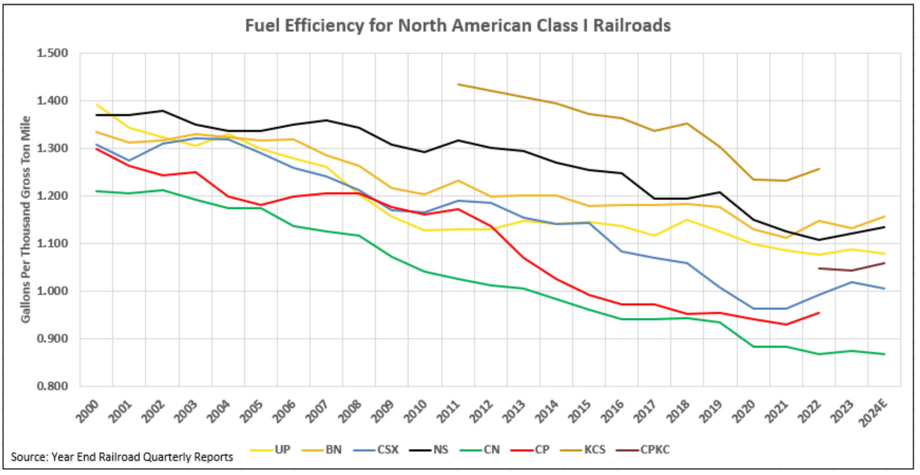


Figure 4 – Fuel Efficiency for Class 1 Railroads

Figure 5 shows the weighted average performance of the Class I railroads, so it is an industry performance metric that indicates a 1.0% annual improvement in fuel efficiency going back to 2011. This gradual improvement has been gained largely through:

1. Purchasing newer, more fuel-efficient locomotives
2. Precision Scheduled Railroading (PSR) which has reduced the Horsepower Per Trailing Ton (HPTT) through building longer trains with fewer locomotives pulling them
3. HPTT rules and enforcement, shutting down or idling locomotives on trains when they are not needed for a crew trip
4. Idling reduction technologies and rules such as Automatic Engine Start Stop (AESS)
5. The implementation of Energy Management Systems (EMS)

If we look at the aggregate SBTi goals out to 2030 and the glide slope required for the industry to achieve that goal, it will require from the current year forward, a 6.8% year over year improvement in fuel efficiency to meet the SBTi targets.

Whereas on the current historical glide slope, that would create a 40% percent gap between where railroads need to be and where they likely will be on their current trajectory or historical glide slope. Per the more challenging 1.5 degrees Celsius SBTi goals to be approved in the future as with UP, this will likely mean the slope of the line shown in Figure 4 will become even steeper.

Note that the use of biofuels such as biodiesel and renewable diesel which have roughly two thirds less life-cycle CO_2 compared to regular diesel fuel will likely shrink this large gap. For example, if the industry adopted an average B20 blend (20% biofuels) by 2030, that would shrink the 40% gap to roughly 27%. In addition, if traffic volumes continue to shrink at the historical rate of approximately 0.9% per year – given the two largest railroads have absolute goals – that would further shrink the gap closer to 20%. As the trucking industry has much higher CO2 emissions per ton mile than rail the overall impact of this traffic reduction on the climate would be negative, however.

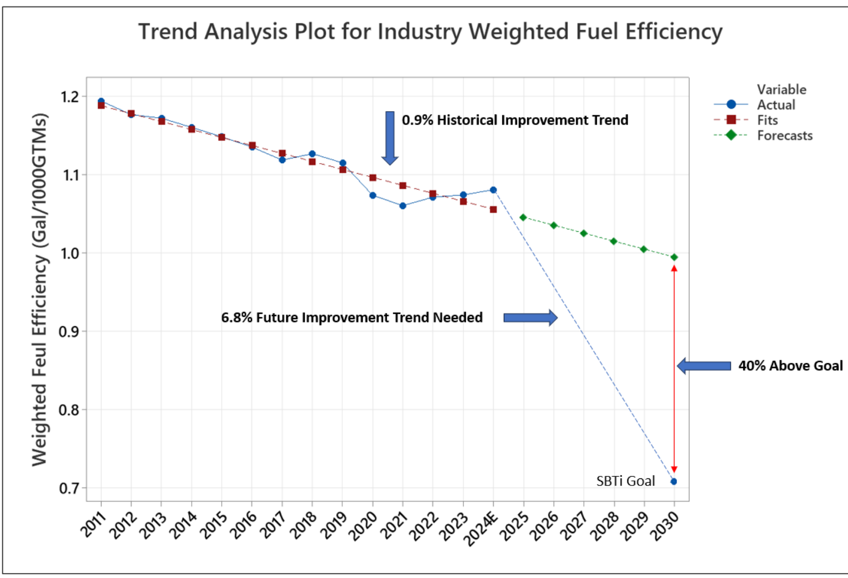


Figure 5 – Historical Rail Industry Glide Slope versus SBTi 2030 Goal

Overview of Available Technologies in Various Stages of Adoption

There are a myriad of technologies or operational approaches available to further reduce railroad fuel consumption. Figure 6 outlines two dozen different technology applications in four general categories:

1. Engine Efficiency
2. Locomotive Power Utilization
3. Train Drag Factors
4. Other

Many of these technologies are in low levels of adoption. There are a variety of reasons for this, but the main culprit is the inherent difficulty in validating small (but important) levels of fuel savings in the 1% to 3% range. There is a large amount of variability in fuel consumption in revenue freight service and the accuracy and reliability of on-board fuel gauges is often problematic. Looking for a 1% to 3% fuel savings when your fuel data is lucky to be within 5% accuracy creates a very large noise to signal ratio. Advanced statistical modeling tools as well as the proper segmentation (grouping based on chosen parameters) of fuel consumption data is required to reliably validate "low level" fuel savings for various technologies or operational approaches.

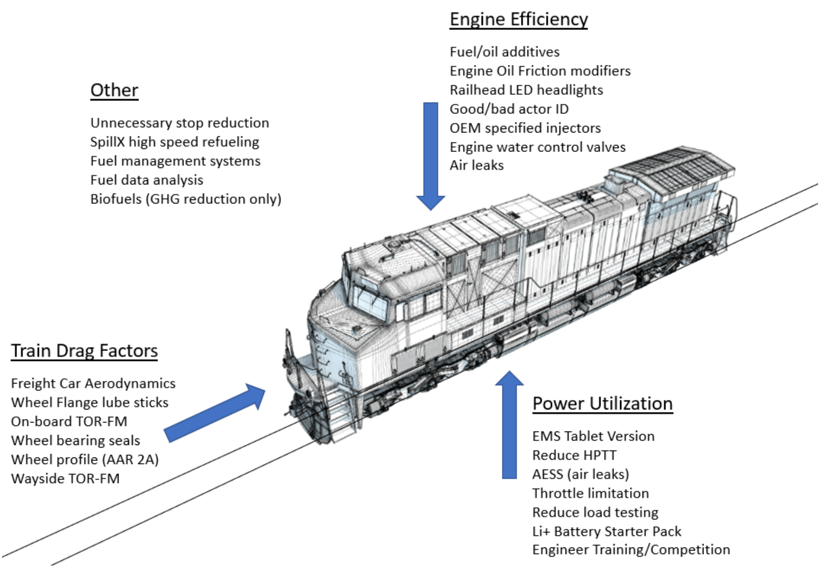


Figure 6 – Various Fuel Savings Technologies and Operational Approaches

VARIABILITY DISCUSSION

Common measures of variability in statistics include range, median absolute deviation (MAD), variance, and standard deviation (sd). You generally want variability to be as small as possible. In a manufacturing world, variability has to be very low or piece parts won't fit together and function as designed or intended.

In the real world of railroading, fuel burn variability is (unfortunately) quite high. There are a lot of reasons for this, and not many are within our ability to control. Below is a partial list of drivers of variability of fuel consumed for a given train:

Variability Drivers

1. Train length
2. Weight or tonnage
3. Train type (coal, intermodal, mixed freight, auto, etc…)
4. Loaded or empty status
5. Horsepower per trailing ton or the number of locomotives pulling the train compared to the total tonnage
6. Topography or the amount and steepness of grade
7. The degree of track curvature
8. Weather conditions such as rain or ice which affect adhesion or traction
9. Wind speed and direction
10. Air pressure due to elevation
11. Ambient temperature
12. Training and skill of the locomotive engineer
13. Age and health of the locomotives pulling the train
14. Age and health of the freight cars on the train
15. Age and health of wheels, trucks and bearings
16. Age and health of air brake system, leakages.
17. Congestion along the route or how often the train stops, either intentionally or otherwise
18. Condition of the track system and underlying ballast
19. The presence of lubrication technologies either wayside or on-board the locomotive or cars
20. Average train speed and associated aerodynamic losses
21. Impact of train makeup and railcar designs on aerodynamics
22. On-board locomotive fuel saving technology such as Energy Management System (EMS)

Additionally measured fuel consumption may not match actual due to

23. Inaccurate fuel gauge
24. Slosh in fuel tank
25. Rail grade impacting fuel tank measurements
26. Incorrect fueling event records
27. Event recorder throttle notch measurements not matching actual fuel burn
28. Wattmeter not measuring fuel burn outside of traction power i.e., not counting accessory loads such as radiator fans, air compressor, etc.

The above list is by no means a comprehensive or complete list of what drives excessive variability in railroad fuel consumption – but we've hit most of the major high points. Due to the large number of variability drivers, it is not uncommon for what is considered the "same" train to burn significantly less or significantly more fuel than the average for that train type and location. Train fuel consumption segment burns 40% higher or lower than the average amount are certainly possible under typical operating conditions.

CREW SEGMENT FUEL MEASUREMENT METHODOLOGY

Accurate fuel consumption data for each locomotive on a train is the crucial input needed to calculate fuel burn at a train level. Segmenting the data by crew segments (crew on to crew off) is a convenient and useful methodology, given crew segment data can be used for many other purposes such as grading locomotive engineer performance and training initiatives or operational changes that may be geographically constrained.

Back to our discussion on sources of variability, there are some sources that can and should be accounted for and which determine how data is gathered, compiled, and segmented to be used for future fuel analysis of any kind. There are a few discrete measures that need to be addressed, such as:

1. Number of locomotives pulling the train - this assumes there may be some locomotives shutdown (DIC or Dead-in-Consist) or idling
2. Type of train (coal, intermodal, mixed freight, auto, etc…)
3. Train length
4. Tonnage
5. Geographic location, or crew point A to crew point B which must include direction as well

Fuel consumption can be obtained from changes in the fuel gauge after accounting for fueling events, the event recorder of throttle notch information, and the on-board Wattmeter. A weighted average of these variables will typically provide the most accurate measurement of fuel consumption when combined with careful error handling to account for locomotives where information is missing or inaccurate. Throttle information is typically the most accurate so it should be weighted the most heavily, however a weighted average of all three will typically be more accurate than throttle information by itself.

The proper dependent variable to use in analysis of fuel consumption technologies is usually what different railroads refer to as either the fuel efficiency (FE) or c-rate and is calculated as 1000*Gallons/(Tons*Miles). This was also described above for Figure 4. Most technologies can be assumed to save more fuel linearly as tonnage and miles increase, so looking at impacts on raw fuel burned will give biased results. The natural logarithm of FE may also be the better variable to use in some cases compared to raw FE as its distribution can be right skewed and it is sometimes more useful to describe treatment impacts as a percentage. Figure 7 below shows the distribution of FE for a given Automotive train crew route on a Class I railroad, there are quite a few outliers in the data. Trains with higher tonnage generally had better fuel efficiency but the data shows a great deal of noise.

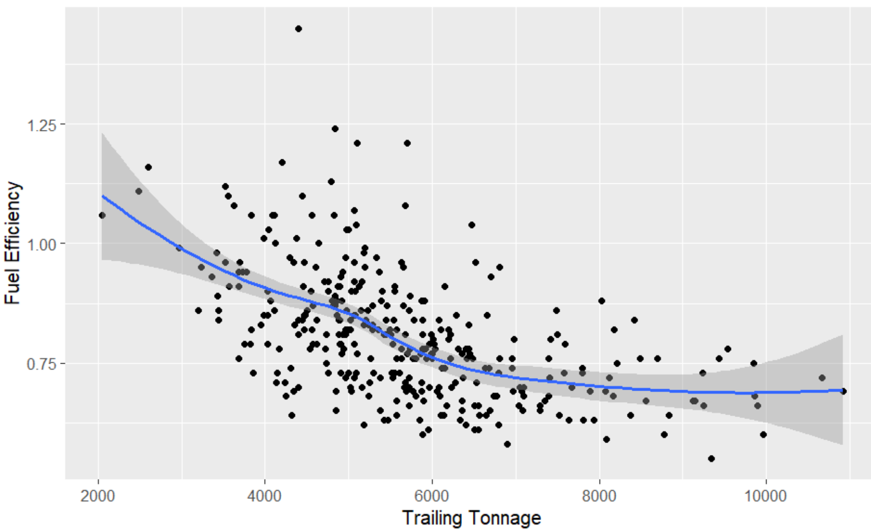


Figure 7 – Representative Regression Graph of Tonnage and FE (gallons per thousand gross ton miles)

Once data has been collected and FE has been calculated, a full regression should be run using a variety of data points as explanatory variables to increase statistical power (probability of obtaining a true positive result) and account for potential sources of bias. It is critical to remember that the results of most field tests should be treated observational studies as fuel saving treatments are often not randomly applied to trains. For instance, many railroads have made a rule that if an EMS equipped locomotive is on a train it should be used as the lead locomotive to run the train. As longer trains will therefore be more likely to have an EMS equipped locomotive with more total locomotives, and as longer trains generally have better/lower FE due to lower HPTT and lower aerodynamic drag per ton (only one front of the train) this will create a source of bias in any study of EMS technologies that does not include tonnage and HPTT as predictors of FE. A list of potential covariates that can be included in a model is shown below. Models should generally include only one rail category as all coefficient estimates will likely be different between Manifest and Intermodal etc. Additional variables which will be useful if obtainable are what % of the locomotives are Tier 4, average locomotive age, average wind speed on the day of the trip, and engineer as a random effect.

Predictor	Type	Explanation
HPTT	Fixed Effect	Locomotives are generally more efficient in higher throttle notches, so fewer locomotives per ton often means better efficiency
Length	Fixed Effect	Longer trains are more efficient, only one front of train
Length/Tons	Fixed Effect	A rough measure of aerodynamic drag per ton, also an impact on wheel/rail friction
Season	Fixed Effect	Can be handled using terms for month or sine(Time) and cos(Time)
Time trend	Fixed Effect	Change over time
Route	Random Effect	There will typically be many routes in a data set so information can be shared between them using a random effect model structure.
Train	Random Effect	A typical train goes through several crew trips over a route. Usually identified using train symbol, train day and train section.

Table 2-Explanatory variables for FE

The table below illustrates how different modeling methods can impact the power of a study (probability of obtaining a true positive result) along with the false positive rate (probability of declaring a statistically significant result at the 0.05 level when there is no true effect). These results were obtained from real railroad fuel consumption data from a Class 1 railroad, with a simulated treatment being applied to half of the crew trips and then repeated using bootstrapping (randomly sampling from the full data set with replacement with a treatment randomly applied to half of the sample) to test the accuracy of different modeling methods. Bootstrapping allows one to test the accuracy and unbiasedness of statistical techniques on real data instead of fully simulated data in addition to other uses.

As seen below in Table 3, the mixed effects regressions on FE or logged FE greatly outperform the others, with an accurate false positive rate of ~0.05. Mixed effects regressions allow for random effects such as train in the model structure, because a typical treatment is applied to trains and not individual crew trips a model that doesn't include train as a variable or uses it as a fixed effect will break the assumptions of linear regression and give a very high false positive rate of around 0.09 as shown below. This means that when there is no true effect the model has a 9% chance of estimating a "statistically significant" difference with a p-value less than 0.05. Not adjusting for any covariates leads to greatly reduced power as shown in the results from T-tests with a true positive rate or power less than 0.2. The True SE or true standard error in the table measures how much noise a model's estimate of the treatment effect has, the lower the better, while the Model SE measures how well a model is able to calculate its own noise. The Model SE should be very close to the True SE, otherwise there is an isse with the method being applied to the data.

Model	False Positive Rate	True SE of Effect Size	Model SE of Effect Size	True Positive Rate (Power)
Ideal Model	0.05	0	=True SE	1.0
T-Test	0.028	4.35	-	0.19
T-Test log(FE)	0.038	5.47	-	0.151
Regression	0.095	2.33	2.01	0.686
Regression log(FE)	0.082	2.59	2.25	0.606
Mixed effects regression	0.054	2.32	2.23	0.588
Mixed effects regression log(FE)	0.055	2.70	2.57	0.507

Table 3-Performance of Different Statistical Methodologies

As fuel data also typically contains outliers, different outlier handling methods can also be used to fix the issues that they can create. Table 4 shows how two different methods for handling outliers, trimming where the outliers more than 2 standard deviations (sd) from expected values are deleted, and censoring where outliers more than 2 sd from expected values are replaced with the expected value plus or minus 2 sd depending on its location. As seen in Table 4, censoring greatly outperforms trimming which inflates the false positive rate due to a model SE that is biased low because of the removal of a great deal of variation in the data. In general, blindly deleting outliers or outlier trimming should be avoided whenever possible as they are often at least partly due to real factors in the underlying data, and methods such as censoring or the use of extreme value distributions should be strongly preferred. Outliers can also be a source of knowledge for operational improvement if the causes for very low or high values can be determined, knowledge that will be discarded if outliers are deleted prior to modeling.

Model	False Positive Rate	True SE	Model SE	True Positive Rate
Ideal Model	0.05	0	=True SE	1.0
Original Model	0.042	2.61	2.65	0.49
Censored Model	0.054	2.51	2.48	0.536
Trimmed Model	0.071	2.52	2.22	0.615

Table 4-Performance of Different Outlier Handling Methodologies

Too often, a technology is tested using an unstructured and haphazard approach, perhaps for a week or a month with no plan as to how the fuel savings will be validated. At the end of the test, if no fuel savings are found, it doesn't mean that fuel savings don't exist, it may mean that they are just being masked by a poorly structured test plan and a lack of statistical rigor in approaching the validation of a certain technology or operational change.

Dedicated Rail Test Facilities

Pueblo Colorado is home to two dedicated facilities for railroad testing. The Transportation Technology Center or TTC[2] (Formerly TTCI) operated by Ensco for the Federal Railroad Administration (FRA) and the Department of Transportation (DOT) and the relatively new MxV Rail[3] facility supported by the Association of American Railroads (AAR).

These facilities offer a unique opportunity to test fuel saving technologies without the inherent noise of revenue freight fuel consumption variability, where train operations are constrained by the necessities of running a railroad. Different track loops of varying track length and curvature with known elevations and grades are available with dedicated test trains measuring precise fuel consumption and drawbar coupler forces which equate to the amount of friction present while pulling the test train.

Testing energy savings technology and aerodynamic treatments on locomotives and rail cars provide a controlled window which can assess the general capability of a given fuel or energy conservation technology. This can be used as a floor once testing progresses to revenue freight service even with the inherent variability of regular train operations.

Case Study: Locomotive Wheel Flage Stick Lubrication

Presented below is a case study or example of a systematic, credible, and carefully designed and executed validation test program of a fuel conservation technology; (locomotive wheel flange stick lubrication).

There have been many attempts to develop formulations to provide lubrication to rolling/sliding elements such as wheel and rail contact in railroads. These formulations range from liquid and grease systems, which generally require more expensive application equipment, frequent monitoring and suffer from plugging applicators which limit effectiveness and reliability, to solid stick formulation which can be applied directly to the wheels or the rails.

These solid stick formulations are generally comprised of some type of binder material and a range of various lubricants. The binder material holds the lubricating materials in place and typically dictates the rates at which these lubricants are applied to the wheel or rail. An extremely hard or wear resistant binder material limits the amount of lubricant applied to the steel surface and a soft material will allow much more lubricant to be transferred to the steel surface.

The goal of applying lubrication is to reduce friction and thereby reduce energy consumption. In order to determine the optimal solid lubricant binder strength needed, testing on actual trains was required to determine the amount of lubrication required to have a measurable fuel consumption reduction benefit. Early testing provided a range of binder strength required to provide the optimal friction reduction benefits between the wheel and rail interface.

In an effort to provide a truly environmentally friendly lubricant, researchers at the Kansas State University Technology Development Institute began investigating a wide range of biopolymers that could be used as a binder material and the possibility of incorporating a vegetable based oil to add friction reduction and improve transfer of the lubricant from the wheel of the locomotives where it was being applied down to the rail, and ultimately onto subsequent wheels of the train providing friction reduction for all wheels of the train adding to fuel savings. Using Polyethylene as baseline for strength and hardness, all commercial biopolymers were investigated to identify a biodegradable and renewable binder material that could be incorporated into the wheel/rail lubrication process.

A range of vegetable oils were also investigated to determine which oil provided the best lubrication benefits. The Tribology Handbook was consulted to determine which vegetable-based oil provided the greatest wear reduction using the standardized 4 ball wear test. Castor oil provides the highest lubrication benefit of all vegetable oils tested.

Once a range of biopolymers had been identified and castor oil selected as the oil of choice, over 200 samples of various blends were created and tested for strength and lubrication efficiency at the K-State labs. This testing narrowed down the formulation and lead to the discovery of a blend of multiple biopolymers that provided the best application rate of material.

Once the formulation had been optimized, production scale-up enabled sticks to be produced that could be installed in locomotive applicators and tested at the TTCI facility on locomotives to determine the coefficient of friction (CoF) reduction on the gauge face of the rail and the energy savings provided by the lubricant.

Analysis of Tests Undertaken at Test Facilities

As mentioned previously, too often, a technology is tested using an unstructured and haphazard approach, perhaps for a week or a month with no plan as to how the fuel savings will be validated. Performing a test in dedicated rail test facility brings structure, yet the data still requires statistical rigor to confidently establish or quantify the fuel savings of a technology.

The following discussion describes such an analysis performed on two separate tests of a fuel saving technology. The tests were run at TTC^2 (Formerly TTCI), in 2012 and 2014. Each test consisted of baseline, or "dry" laps, as well as "lube" laps where an environmentally friendly solid polymer friction modifier formulation was applied.

The objective of the study was to quantify the difference between the lubricated versus dry conditions. Statistical analysis brought rigor into the tests to (1) provide a formal statistical test as to whether there were any effect of the claimed energy or fuel savings benefits and (2) if there were a benefit, provide a statistically determined estimate of its likely range (versus a point estimate).

Tests were conducted at TTCI in 2012 on the Wheel Rail Mechanism (WRM) loop and in 2014 on the Transit Test Track (TTT). These tracks have different profiles with respect to curves, and to some extent, elevation changes. Figure 8 depicts the tracks and aspects of the tests.

Methods for analysing data from track tests with proper treatment designs can often be much simpler than those needed for revenue testing, t-tests with no covariates or outlier handling can be appropriate when a test is fully balanced with all other variables such as engineer behavior, weather, and train build held constant.

The results of this testing indicated that within a short time period of application, energy savings provided by the lubricant ranged between 2% and 4.5% for both mechanical and electrical energy and the material was able to immediately reduce the outside rail gage face CoF from 0.44 down to 0.26 based on tribometer measurements in a single pass from 2 locomotives and 30 loaded hopper cars.

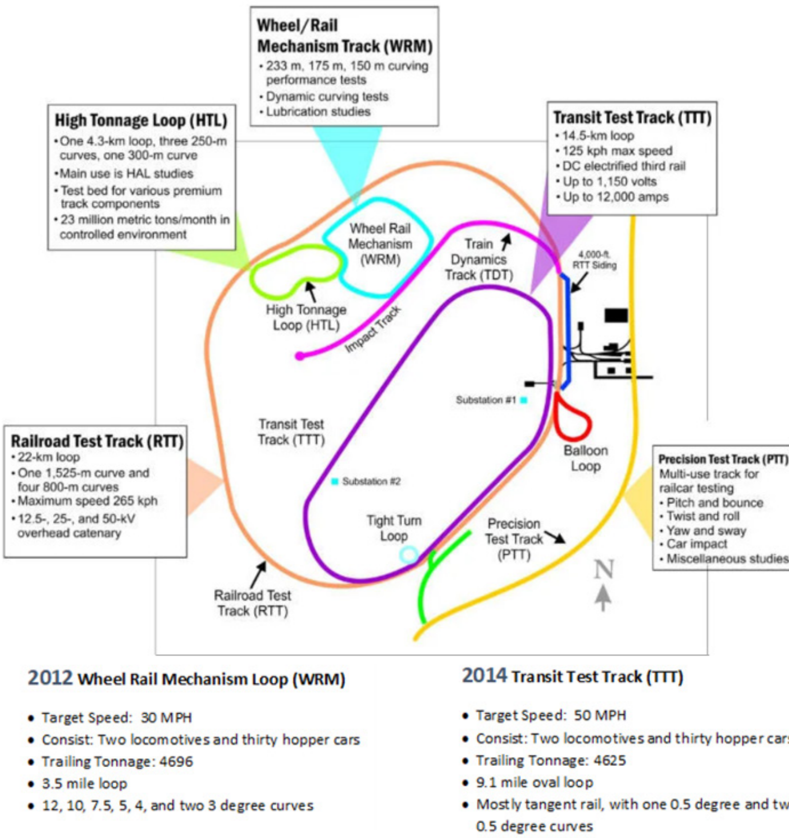


Figure 8 – TTCI Test Track Configurations[2]

The locomotives were equipped with wattmeters for traction power as well as a secondary drawbar mechanical coupler. From these two sources we were able to compute both electrical and mechanical work, in kilowatt hours (KWHR). Our target metric for estimating energy savings is electrical energy in KWHR. A check was performed to ensure that electrical and mechanical work were closely correlated, to build confidence in the measurement systems. The correlation coefficients were 0.9999 for the 2012 WRM test, and 0.9996 for the 2014 TTT test.

A GPS system provided the location (latitude and longitude) of the locomotive on the track. This allows us to augment the core energy data with:

- **Train speed**. Speed does vary slightly from the target speed, especially on the WRM which has more curves and grade changes.

- **Heading**. This allows us to examine potential correlations of curvature with power consumption.
- **Elevation**. Latitude and Longitude information was used to obtain elevation information from government sources, allowing us to examine potential correlations of grade with power consumption.
- **Mileage Traveled**. Though this may not be a modeled measure, it does help us in the computation of measures such as speed and grade change.

Measurements were recorded at one-second level (one-hertz data). Information about the lap number and the lubricating condition (lube or dry) was provided. Some filtering was applied to exclude records (laps or portions of laps) that were designated as conditioning laps. Conditioning laps have the purpose of drying up and dispersing any residual grease from prior tests such that there is a dry coefficient of friction as the new test begins. The goal in filtering the data was to have a clear, contrasting picture of the lubricating versus the dry condition.

When a single treatment condition B is compared to a control condition A using time series data as at TTCI, it is generally best to format the order of tests where possible as A-B-B-A, referred to as an ABBA treatment design, where an A refers to a loop run or set of loop runs with no lubricant present and B refers to an equal set of loops with the lubricant applied. When an AABB treatment design is used instead, the impact of the treatment will become correlated with any time trends that are present almost every day due to weather or other factors, and this could heavily bias the results.

The data was combined across the two tests. A variety of statistical tools were applied to appropriately analyze the data generated by the runs and ensure that the results could be projected to real world contexts. For more information, see the full report[4].

Figure 9 shows the overall fuel savings due to the wheel flange lubrication sticks is estimated to be 3.2%. This represents a statistically adjusted estimate of the percent savings difference between the lubrication and Base runs of the combined TTCI tests. The results are statistically adjusted in the sense described in the methodology section. Analysis of Covariance (ANCOVA)[5] methods quantify potential other causal factors, such as speed, curvature, and elevation changes, to balance the comparison and isolate the effect due solely to the lubricant. The raw mean difference between the groups was 3.3%. The fact that this was a very modest adjustment indicates that the tests were well-executed to minimize any differences that might mistakenly be attributed to the lubricant.

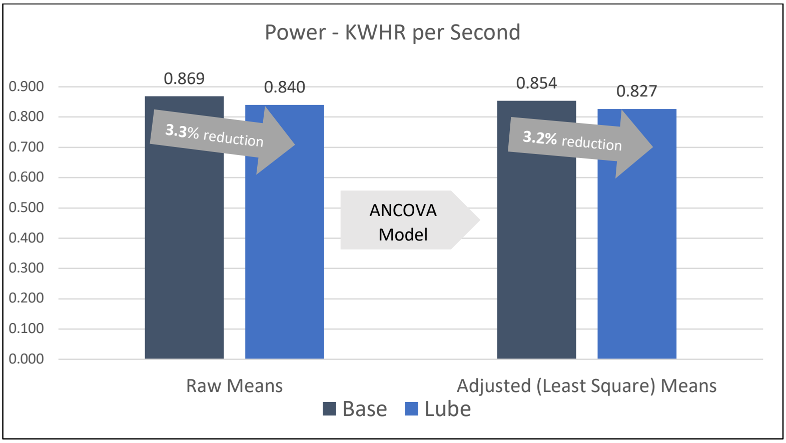
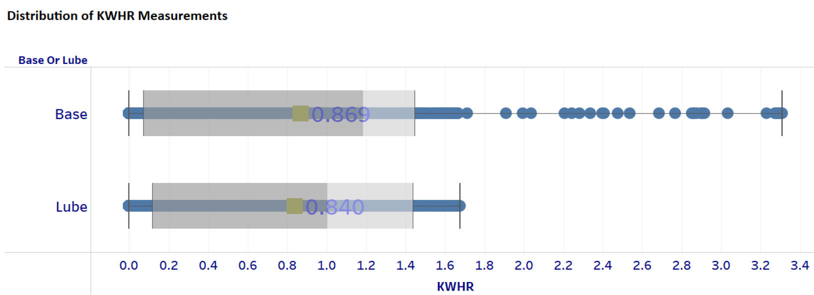


Figure 9 – Comparison of Means, Base versus Lubricated

Because of the very small adjustments to the means, we can, in fact, look at the dispersion of the raw, unadjusted data, to get some insight into the differences. Figure 10 is a box plot showing the spread of the measurements for the Base versus Lube conditions. In the case of this visualization, the green box represents the mean, with the mean KWHR for Base being slightly higher than that for Lube.

We also see that for the Base condition there are several measurements of high energy expenditure (the points above 2.0). Although one may label these as outliers and conjecture that these alone are driving the difference in means, in reality, these are a small number of observations among 8681 total observations for the Base runs. Furthermore, the ANCOVA approach allows us to study this variation as a whole in a way that accounts for these kinds of events.

Base skews slightly higher than Lube with higher measurements



Condition	Number of Observations	Mean	Standard Deviation
Base	8,681	0.869	0.659
Lube	11955	0.840	0.636

Standard Deviation is a measure of spread. Base has a greater range of energy measurements values.

Figure 10 – Distribution of KWHR Measurements

An output of the ANCOVA model is a statistical estimate of the hypothesis that the effect of the lubricant is different from zero[6]. In statistics parlance, this is the so-called "null hypothesis" – that of no difference between two measured phenomena[7].

Our formal statistical test indicates that we can reject the null hypothesis of no effect and can do so at the greater than 99% confidence level (t-value of 3.56). That is to say, there is a very small chance that there is no effect of the lubricant.

A second, but less generalized test of differences is the so-called two-sample t-Test[8]. It is less powerful and accurate than ANCOVA with covariates in that it is a univariate approach comparing the simple means and variances, without accounting for other factors, such as those described previously, that could influence those means and variances. It is only presented here to give the reader an understanding of the raw data. Again, the null hypothesis of no difference between the means was rejected at the greater than 99% confidence level (t-value of 3.14 using the Satterthwaite test of unequal variances).

In addition to these tests, we can use the same model to provide an estimate of the range of possible effects, versus a single point estimate such as the mean. This "histogram"[9] provides an extra dimension to our understanding of the effect.

Figure 11 shows how the ANCOVA model simulates the spread of the possible effect, where the height of the bars indicate the relative probability of the effect being in the range shown in the horizontal axis. In other words, if we were to conduct the tests many thousands of times, the effects would fall into these “buckets.”

The 90% confidence interval for fuel savings is between 1.8% and 4.3%, the 90% confidence interval will capture the true parameter value 90% of the time when a statistical test is run.

It is worth noting that the histogram does not cross zero. That is to say, it is very improbable that the lubricant has no effect absent some unknown bias in the testing.

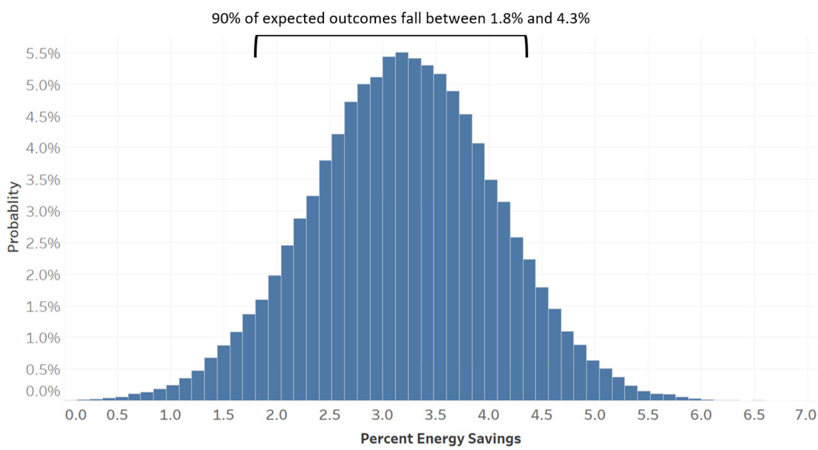


Figure 11 – Range of Estimated Savings

Other analysis, not included in this paper, were also performed. You can download the full report for details.[10]

To provide confidence in the test results the analysis approached the test data from a number of angles. A battery of statistical tools, including models and formal statistical tests, were applied. The following table provides a summary of those analyses.

Analysis Metric	Test or Methodology	Result
Energy Savings	ANCOVA - Estimation	Savings estimated to be 3.2%
Energy Savings	ANCOVA– Hypothesis Test (t-test)	“No effect” hypothesis rejected at 99% confidence level
Energy Savings	Simple two-sample t-test	“No effect” hypothesis rejected at 99% confidence level
Energy Savings	Confidence Intervals	90% of expected outcomes are within 1.8% and 4.3%
Throttle Position	Simple statistics and visualization	7.1% less time spent in T8 (not statistically adjusted) More time spent in T4 and T5
Throttle Position	M-H Chi-Square test	Time spent in each throttle position, taken as a whole, is statistically different at the 90.1% confidence level
Throttle Position	Simple two-sample t-test	Average throttle position is statistically different at the 90.0% confidence level
Throttle Position	Logistic Regression - Estimation	Odds of being in T8 reduced by 4.8%
Throttle Position	Logistic Regression – Hypothesis Test (Wald Chi-Square)	“No effect” hypothesis on T8 reduction rejected at 87.1% confidence level.
Throttle Position	Visualizations	Visible gap between Lube and Base along the track mileage

Table 5 – Summary of Various Methodologies Applied

Leveraging the data from the two TTCI tests (2012 and 2014), a third study produced a model to project the results from the tests to real-world field contexts[11]. This allows a railroad to input the attributes of their network and operations, such as grade profile, tangent versus curve percentage, typical locomotive counts, train length, and the application protocol for the lubricant. Figure 12 shows the output of the end-user tool for a typical Western US railroad.

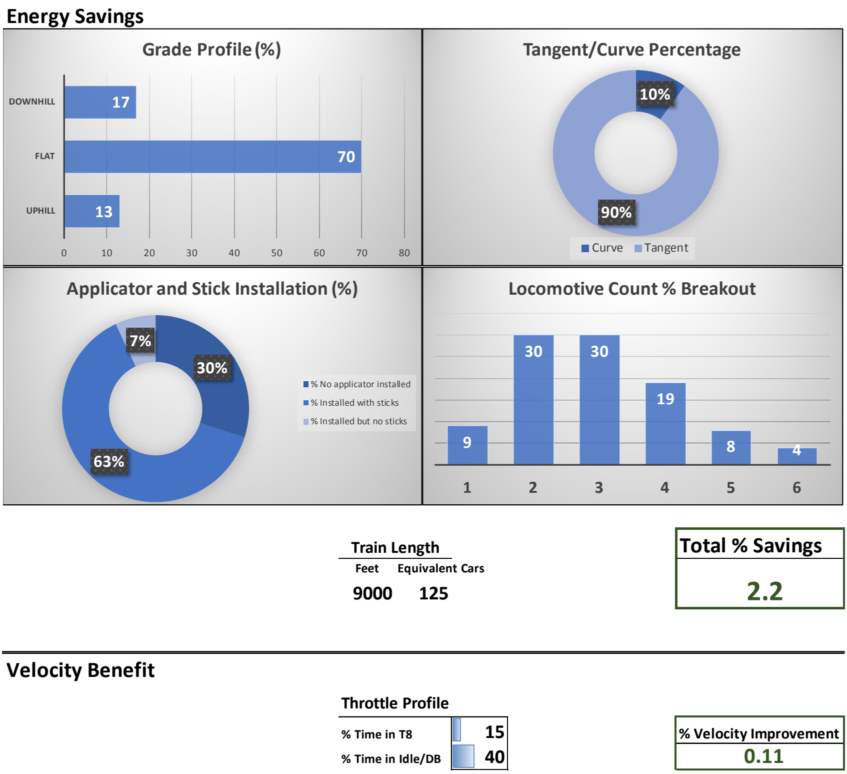


Figure 12 – Projection of Energy Savings Based on Various Operating and Track Parameters

Revenue Field Testing at a Class I Railroad

Revenue testing was performed using these products on a class I railroad from 2014 through 2015 with data including over 500,000 individual crew trips. Both log(FE) and trip velocity in MPH were modeled using the mixed effects regression method described above. Separate models were developed for Intermodal, Manifest and Coal train categories and combined using Meta-analysis. Meta-analysis is commonly used to combine several medical studies to calculate a single overall average treatment effect. This analysis found an average savings from flange sticks of 1.4% with a 95% confidence interval between 0.7% and 2.2% when one locomotive had stick brackets equipped at the front of the train during the study period. When multiple locomotives had brackets equipped the savings increased to 1.8% with a 95% confidence interval of between 0.6% and 2.9%. Brackets also increased speeds by 0.09 MPH but this impact was not statistically significant. A differences-in-differences study design utilizing both bracket presence and study start date as commonly seen in Econometrics was used to identify causal impacts from flange stick usage.

Conclusions

Accurately measuring the fuel savings of many different locomotive technologies and rail operational changes (lowering HPTT as an example), as well as various locomotive engineer training initiatives is challenging. Given the vast variability of day-to-day operations for any given Class I railroad and somewhat unreliable on-board locomotive fuel gauges, the problem becomes compounded quickly.

A rigorous statistical approach to this problem can yield results, usually with a good degree of precision and a high level of confidence.

When investigating new technologies in particular, available test facilities such as TTC and MxV Rail are excellent resources to perform energy testing. If Class I's decide to pool resources and agree on a structured test regimen, the cost considerations can be very reasonable while providing controlled results that all railroads can use as a baseline.

For railroads that may lack the expertise to design, build and perform the statistical tests outlined in this paper, there are statistical companies and consultants who have a specialty and rich history of performing this type of work for railroads. Use them in conjunction with testing performed at TTCI or MxV Rail and revenue field testing also.

Revenue field testing is the logical next step as there likely will be variations on actual savings from one railroad to the next, driven by differences in HPTT, commodity mix, track topography and other factors. When designing a statistical test, keep in mind the following factors:

1. Allow enough time to test the technology completely, this is usually measured in months
2. Test through a broad range of scenarios across the network and across seasons as well
3. Compare baseline to test measurements concurrently so as to avoid seasonal issues
4. Design a balanced model which can be analyzed several different ways with various statistical comparison tools

Avoid the mentality trap of "we couldn't prove anything within a matter of weeks, so the technology must not be saving any fuel". Generally, most technologies do provide some level of fuel savings; the goal is to accurately determine the appropriate range of savings in order to calculate a reasonable return on investment or ROI.

Another concern is succumbing to a "groupthink" where a certain technology falls out of favor for no good reason at one railroad and others follow suit, not driven by any firm data analysis.

The number of fuel-saving technologies is varied and plentiful – many of them are in low levels of adoption due to an inability to effectively measure fuel savings. The tools and techniques outlined in this paper can help railroads as they work to reduce their GHG emissions to meet their short term SBTi goals.

REFERENCES

1. https://sciencebasedtargets.org/news/sbti-raises-the-bar-to-1-5-c
2. https://ttc-ensco.com/
3. https://www.mxvrail.com/
4. https://firstanalytics.com/wp-content/uploads/NatureBlend-TTCI-tests-analysis.pdf
5. https://en.wikipedia.org/wiki/Analysis_of_variance
6. https://en.wikipedia.org/wiki/Statistical_hypothesis_testing
7. https://en.wikipedia.org/wiki/Null_hypothesis
8. https://en.wikipedia.org/wiki/Student%27s_t-test
9. https://en.wikipedia.org/wiki/Histogram
10. https://firstanalytics.com/wp-content/uploads/NatureBlend-TTCI-tests-analysis.pdf
11. https://firstanalytics.com/wp-content/uploads/NatureBlend-TTCI-projection-model.pdf