CIAMTIS

U.S. DOT Region 3 University Transportation Center

Railroad Track Performance Monitoring by Advanced Sensor Network and Big Data

February 28, 2021

Prepared by:

H. Huang and A. M. Zarembski

r3utc.psu.edu

PennState College of Engineering

LARSON TRANSPORTATION INSTITUTE

DISCLAIMER

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CIAM-UTC-REG11		

4. Title and Subtitle	5. Report Date
Railroad Track Performance Monitoring by Advanced Sensor Network and Big Data	February 28, 2021
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Hai Huang, Ph.D., P.E., Penn State Altoona https://orcid.org/0000-0002-0023-2915 and Allan M. Zarembski, P.E., FASME, Hon. Mbr. AREMA, University of Delaware https://orcid.org/0000-0002-4282-9330	LTI 2022-06

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Penn State Altoona 3000 Ivyside Park Altoona, PA 16601	11. Contract or Grant No. 69A3551847103

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
U.S. Department of Transportation Research and Innovative Technology Administration 3rd Fl, East Bldg E33-461 1200 New Jersey Ave, SE Washington, DC 20590	Final Report 3-1-2019 – 9-30-2021
	14. Sponsoring Agency Code

15. Supplementary Notes

Work funded through The Pennsylvania State University via the University Transportation Center Grant Agreement, Grant No. 69A3551847103.

16. Abstract

Long-lasting track quality is dependent on a well-maintained track ballast and subgrade. Accelerated deterioration of ballast and sub-ballast layers ultimately leads to accelerated deterioration of the track superstructure that rests upon it. It has been shown through research and practice that minimizing variations in track geometry variables is the best way to guarantee safe and efficient track system usage. This report focuses on the development of analysis techniques for data collected from test sites containing a total of 16 SmartRocks. SmartRocks are a new measurement hardware designed with maintenance engineers in mind, capable of measuring ballast dynamics. The analysis makes use of data science techniques to identify differences in SmartRock measurements to quantify ballast response. The use of SmartRocks provides measurements of ballast dynamics under passing trains, quantifying inter-particle interactions in a way that other ballast assessments cannot match. The analysis presented herein provides an in-depth look at SmartRock behavior using the time series' data measured under train passings and how differences in their behavior correlate with measured SmartRock data. This report concludes with a discussion of the results obtained and their potential implications for future ballast and SmartRock research.

17. Key Words	18. Distribution Statement
Railroad, track ballast, particle geometry, superstructure, SmartRock, accelerated deterioration	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	81	

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Long-lasting track quality is dependent on a well-maintained track ballast and subgrade. Accelerated deterioration of ballast and sub-ballast layers ultimately leads to accelerated deterioration of the track superstructure that rests upon it (Li et al., 2016). Avoiding severe degradation in a track section requires well-timed track tamping, undercutting, and/or other maintenance measures such that track geometry does not degrade to critical levels and the track section can remain in service. Federal Track Safety Standards require that track geometry parameters be kept within certain limits for the track section to remain operational, making the assurance of good track geometry standing imperative for railroads (Federal Railroad Administration, 2008). Ideally, maintenance is performed efficiently and during times where no trains are scheduled to pass over the track section such that service is never disrupted.

It has been shown through research and practice that minimizing variations in track geometry variables is the best way to guarantee safe and efficient track system usage (Hamid et al., 1983). Track geometry defects, defined as large variations from geometry parameters' design values, are a leading cause of derailments, the leading type of all rail accidents. In 2019, 60.4% of train accidents reported by the Federal Railroad Administration (FRA) were derailments, as shown in Table 1 (Federal Railroad Administration, 2019). Maintaining a section of track such that it does not permanently and differentially displace is a primary way of reducing derailments. For this reason, track geometry is a primary area of inspection to maintain good track condition and operations. Correcting track geometry defects is typically done by tamping or undercutting the track (Zarembski & Bonaventura, 2010).

Table 1. FRA data on train accidents by type and cause.

	Total	Reportable Damage ($)	Casualty		Causes					
	Cnt	Amount	Kld	Inj	Eqp	HRC	Hmn	Othr	Sig	Trk
Derailments	1,325	242,106,776	0	21	176	-	467	161	20	501
Head on collision	2	149,612	0	0	-	-	2	-	-	-
Rear end collision	12	5,595,535	0	4	-	-	12	-	-	-
Side collision	69	10,343,385	0	5	1	-	56	8	3	1
Raking collision	26	1,566,688	0	1	3	-	18	1	4	-
Broken train collision	5	1,440,073	0	2	2	-	2	-	-	1
Highway-rail impact	227	28,481,500	61	200	-	227	-	-	-	-
Obstruction impact	81	6,280,916	0	17	-	-	18	60	-	3
Fire/violent rupture	57	9,617,873	0	0	35	-	-	21	-	1
Other impacts	211	15,144,329	0	5	22	-	143	31	13	2
Other events	178	11,065,868	3	2	36	-	82	42	3	15
-- Total	2,193	331,792,555	64	257	275	227	800	324	43	524

Track geometry defects typically occur due to degradation in the trackbed layers (Zarembski et al., 2018; Yurlov et al., 2019; Zarembski et al., 2020). Figure 1 illustrates a cut-section of a typical section of ballasted track with trackbed layers noted. The subgrade is the foundation layer of ballasted track systems but is easily subject to deformation due to the large forces imposed by dynamic wheel impacts. Passing trains deliver forces through the track to the ties, which are then distributed throughout the ballast layer, mitigating as much of the force as possible before it reaches the subgrade. Maintenance of the ballast layer is crucial such that it successfully distributes dynamics forces over its design life, protecting the subgrade, and such that variations in track geometry parameters are minimal.

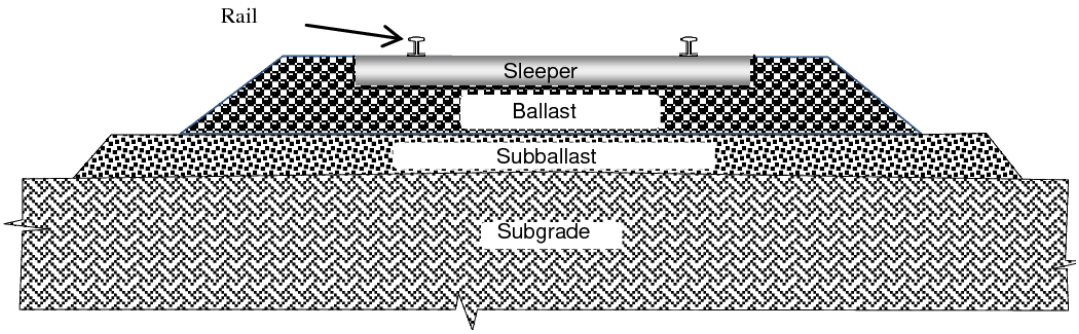


Figure 1. Simplified cut-section illustration of a typical ballasted track system.

Ballast layer strength is correlated with ballast particle shape, angularity, and texture. In particular, ballast particles with more angular shapes interlock with surrounding ballast particles more effectively, increasing the layer's overall strength. In general, well-proportioned particles with a more cubical shape are preferred to particles that are flaky, rounded or smooth (Li et al., 2016). As ballast sections withstand dynamic wheel impacts over long stretches of time, abrasion between ballast particles tends to reduce particle angularity and to break up the particles into smaller grain sizes, ultimately leading to ballast fouling. The amount of ballast fouling is correlated with the percentage of a ballast layer by weight that is made up of tiny particles. Examples of strong and weak sections of ballast are shown in Figure 2, with weaker sections having broader ballast gradation and a larger proportion of fouled ballast material (Li et al., 2016).

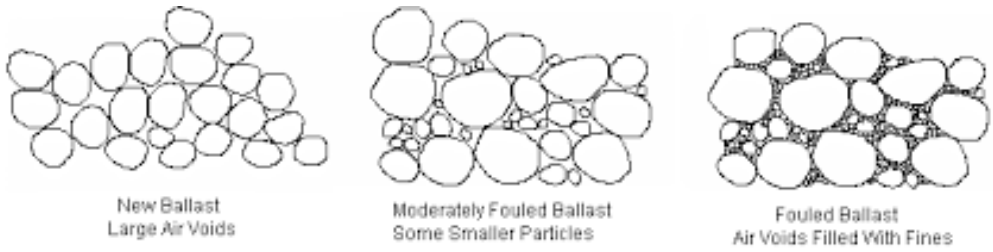


Figure 2. Image comparing a strong, well-proportioned ballast to weak ones.

Degradation occurs in ballast sections at various rates based broadly on the deterioration of desirable ballast characteristics. Following installation of new or rehabilitated ballast sections, maintenance is typically scheduled to occur over a fixed timeframe or scheduled based on reports from manual or automated inspection. With improvements in advanced measurement systems and big data technologies, the railroad engineering industry could benefit from improving its maintenance scheduling procedure by relying more on quantitative information than qualitative information. Quantifying ballast condition and dynamics more accurately would ultimately lead to maintenance being performed more efficiently.

SmartRocks are a new measurement hardware designed with maintenance engineers in mind, capable of measuring ballast dynamics. SmartRocks are wireless sensors that physically resemble ballast particles in shape, size, and texture that are installed in ballast sections (see Figure 3). SmartRocks measure acceleration, orientation, and stress using internal tri-axial accelerometers, tri-axial gyroscopes, and tri-axial magnetometers, respectively. These state vectors are measured in real time, detailing the SmartRocks reaction to passing trains, revealing ballast dynamics unlike the traditional measurement systems used to investigate ballast condition. Acceleration data are measured at a frequency of 50 Hz while orientation data are measured at a frequency of 16 Hz.

Figure 3. SmartRocks compared side-by-side to ballast particles.

This report focuses on the development of analysis techniques for data collected from test sites containing a total of 16 SmartRocks. The analysis makes use of data science techniques to identify differences in SmartRock measurements to quantify ballast response. The use of SmartRocks provides measurements of ballast dynamics under passing trains, quantifying inter-particle interactions in a way that other ballast assessments cannot match. The analysis presented herein provides an in-depth look at SmartRock behavior using the time series data measured under train passings and how differences in their behavior correlate with measured SmartRock data.

Sensitive measurement hardware such as SmartRock provide a new avenue of approach to study ballast response and performance by focusing on inter-particle ballast dynamics rather than simply analyzing track geometry and track deflection measurements. This method of analyzing ballast degradation is relatively new and utilizes advances in data science techniques to explore particle-scale characteristics of ballasted track in an effort to investigate track performance under cyclic loading (Liu et al., 2017). Successful implementation of SmartRock technology allows for a better understanding of ballast dynamics, which may lead to the ability to model ballast deterioration and to schedule proper maintenance interventions to prolong the lifecycle of ballasted track sections. While measurement systems like the Cone Penetrometer Test (Chrismer & LoPresti, 1996) and Ground Penetrating Radar (Sussmann et al., 2001) provide methods of measuring track substructure condition, they require on-rail operations that cannot be performed frequently, thus providing an instantaneous measure of the substructure condition in time rather than a continuous one. Conversely, SmartRocks can record data from every train passing during their installation life, allowing their time series data to be used for investigations into ballast deterioration over time.

This report details all research activities done to clean and analyze recorded SmartRock data. First, previous SmartRock research and the track locations where the SmartRock data detailed in this report were recorded are discussed to provide background on the area of research. Next, the available dataset and all pre-processing steps necessary to extract meaningful time series of passing trains are explored. Exploratory data analysis was performed on the collection of acceleration and Euler angle time series yielded from pre-processing to explore the shape of the data and differences between train passings. From this, descriptive summary statistics were calculated that effectively summarize information about different train passings. Unsupervised machine learning techniques such as K-means clustering, logistic regression, and Fourier analysis are used to quantify differences in the datasets and to determine what characteristics illustrate these differences most accurately in the Results chapter. This report ends with a discussion on the results obtained

and their potential implications for future ballast and SmartRock research. Future areas of improvement and study are suggested as part of the conclusion.

CHAPTER 2

Background

This research project began with a detailed review of all research to date performed using SmartRock sensors. Data have been previously recorded by SmartRocks in laboratory settings (Liu et al., 2017) and in field settings (Liu et al., 2019) to analyze varying track substructure conditions. From this research, several SmartRock analysis techniques were developed, which are utilized in this research. Using state-of-the-art big data analysis techniques, this research project made use of a new SmartRock dataset to analyze and characterize individual SmartRocks based on categorical information to classify ballast response in the track sections in which the SmartRocks were installed.

SmartRocks are advanced, ballast performance monitoring, wireless sensors that are installed in ballast sections. Methods for determining ideal maintenance of ballast sections, such as tamping, undercutting, or ballast cleaning, have traditionally been determined based on professional experience, visual inspection over time, or using automated inspection techniques such as track geometry measurements, whose results are correlated to substructure condition. Since SmartRocks take measurements continuously, they have the potential to identify key behavioral characteristics from which track degradation could be linked. Once these characteristics are identified, the section where the SmartRock is installed can be noted as needing maintenance.

PAST SMARTROCK RESEARCH

Research into responses for SmartRocks installed in track sections with varying conditions and known issues has been conducted over the last decade. Laboratory research of SmartRock responses has been conducted to investigate differences in their response under varying conditions in a ballast box (Liu et al., 2017; Shenton, 1978; Zeng et al., 2019). Ballast boxes are historically used ballast configurations that simulate field ballast conditions by using tri-axial cylinders or simulated ballast layers. Ballast box tests have been used to show the importance of horizontal and rotational SmartRock movement under cyclic loading (Liu et al., 2017) and the sensitivity of SmartRock response to the amount of moisture content in fouled and clean ballast samples (Zeng et al., 2019). These studies validated SmartRock's ability to accurately measure particle vibration, and their conclusions were carefully considered when analyzing the available SmartRock dataset.

Furthermore, site testing of SmartRocks has been performed as well. These studies showed that SmartRock sensors can be used to study ballast dynamics outside of a lab setting. SmartRock acceleration response was notably different when installed in a clean ballast section versus a moist ballast section experiencing mud pumping (Liu et al., 2019). Furthermore, this study showed that SmartRock peak acceleration ranges were correlated with known installation site ballast condition.

This research project aimed to expand upon previously performed SmartRock research in an effort to develop big-data-based analysis techniques that are capable of characterizing ballast condition. The effort into this development, from characterizing all available SmartRock data to identifying measures that correlate with ballast condition, is discussed throughout the remainder of this report.

SMARTROCK DATASET

Each SmartRock is capable of recording time, tri-axial acceleration, orientation (via quaternions), temperature, and tri-axial pressure. SmartRock specifications are shown in Table 2 (Liu et al., 2019). Paired with the SmartRock three-dimensional acceleration time series and four-dimensional quaternion data were a date and time stamp (hereinafter referred to as datetime) and index value (packet number) for each SmartRock instantaneous measurement. Initial exploratory data analysis (EDA) revealed that measurements were not available for every integer packet number of acceleration. Furthermore, quaternion vectors were not available for every packet number for which acceleration vectors were available due to different measurement frequencies. Thus, SmartRock time series were interpolated such that the data reflected a measurement being taken for each integer packet number. This is explained in detail in the Data chapter.

Table 2. SmartRock specifications.

Items	Parameters
Size	6.0 × 6.0 × 6.0 cm
Total weight	295 g
Wireless communication	4.0 BLE Bluetooth
Communication distance	Up to 20 m
Accelerometer	Range up to ±16 g, resolution down to 0.001 mg
Gyroscope	Range up to ±2,000°/s, resolution down to 10^{-5}°/s
Magnetometer	Triple 16-bit, ±4,800 μT
Sampling rate	Selectable data rates up to 512 Hz
Power options	Lipo battery on-board charging via USB
Power consumption	Normal: 10 mA; leep: 0.5 mA
Temperature range	−20 ~ +110°C

Note: USB = Universal Serial Bus.

Data were collected for 16 SmartRocks that were installed in the vicinity of ties in 2 track sections owned by a major North American Class 1 freight railway. SmartRocks were installed at two track sites 2.16 miles apart, one with under-tie pads and one without, as shown in Figure 4. The SmartRocks were installed in the area of turnouts as shown in Figure 5. This figure also shows the server where the SmartRock and strain gauge data were stored.

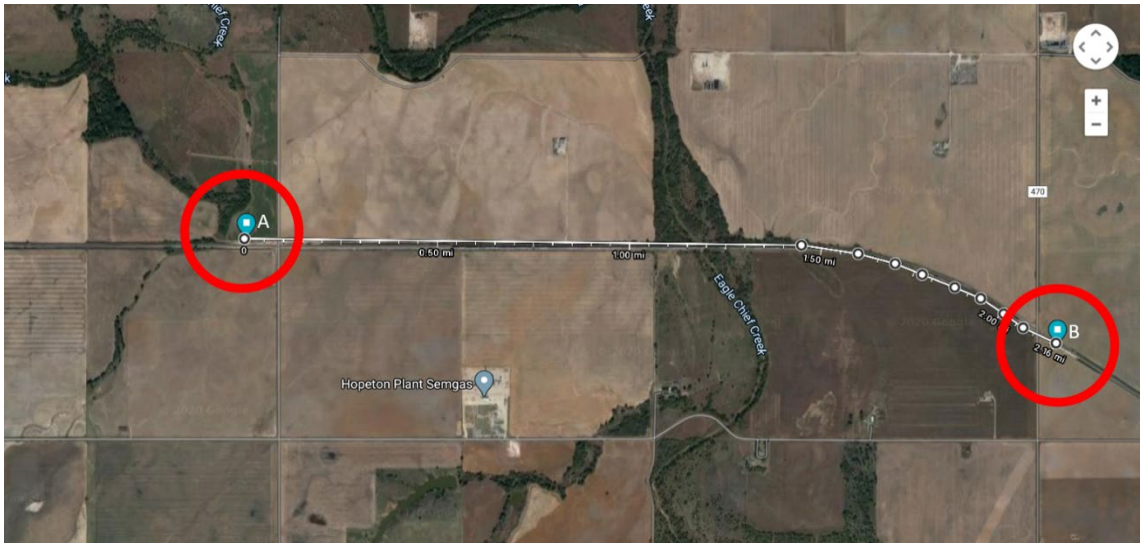


Figure 4. Overview of the track sites where SmartRocks were installed.

Figure 5. Photo showing the track layout where the SmartRocks were installed.

The two sites at which SmartRocks were installed differ in one key area; the Site A track section includes under-tie pads underneath the ties that SmartRocks were installed near, while the Site B track section does not. Under-tie pads are installed to protect the ballast by reducing vibrational forces delivered to the top layer of the ballast. Thus, the effect of this installation should show up in the data, which are discussed in the Results section.

Each site had eight SmartRocks installed, either gauge-side or field-side, and additionally in the ballast cribs (sections of ballast between railroad ties that rests at a height higher than the base of the tie), either in front of the tie or behind the tie (based on incoming traffic). This installation pattern is shown in Figure 6. All SmartRocks were installed 2.5 inches deep in the ballast, thus eliminating differences in SmartRock depth as a potential factor in quantitative differences seen in their acceleration measurements.

SmartRocks are likely to change orientation after installation and are also unlikely to have the same orientation when installed as they did when designed in the laboratory. Thus, all SmartRocks were transformed to a global coordinate basis so analysis across them would be uniform. The internally measured orientation for each individual SmartRock was determined against the uniform track coordinate basis shown in Figure 7. Each SmartRock's three-dimensional acceleration vector was transformed at every time step according to rotation matrices calculated from the SmartRocks orientation data (in the form of quaternions). After the SmartRock acceleration data were transformed into a uniform basis, a secondary, more extensive EDA was performed.

An autonomous and quantitative method of characterizing ballast condition that helps optimize maintenance is greatly needed in railroad engineering and could greatly reduce costs associated with maintenance. Ideal maintenance performance keeps track sections in good standing and reduces costs associated with remedying track sections facing severe deterioration due to poor maintenance scheduling and implementation. Research efforts related to uncovering hidden ballast dynamics using SmartRock hardware are ongoing and hope to provide additional information to maintenance engineers and to provide further quantitative information about ballast dynamics. Studies of track dynamics tend to focus on track geometry and static track substructure measurements for track condition assessments. Quantitative measures on ballast dynamics along with the previously mentioned measurements provide a better, unified system of information for maintenance engineers. This report details analysis methods used to obtain useful information from SmartRock response data and how this information can be used to further understanding of ballast dynamics in track sections where SmartRocks are installed.

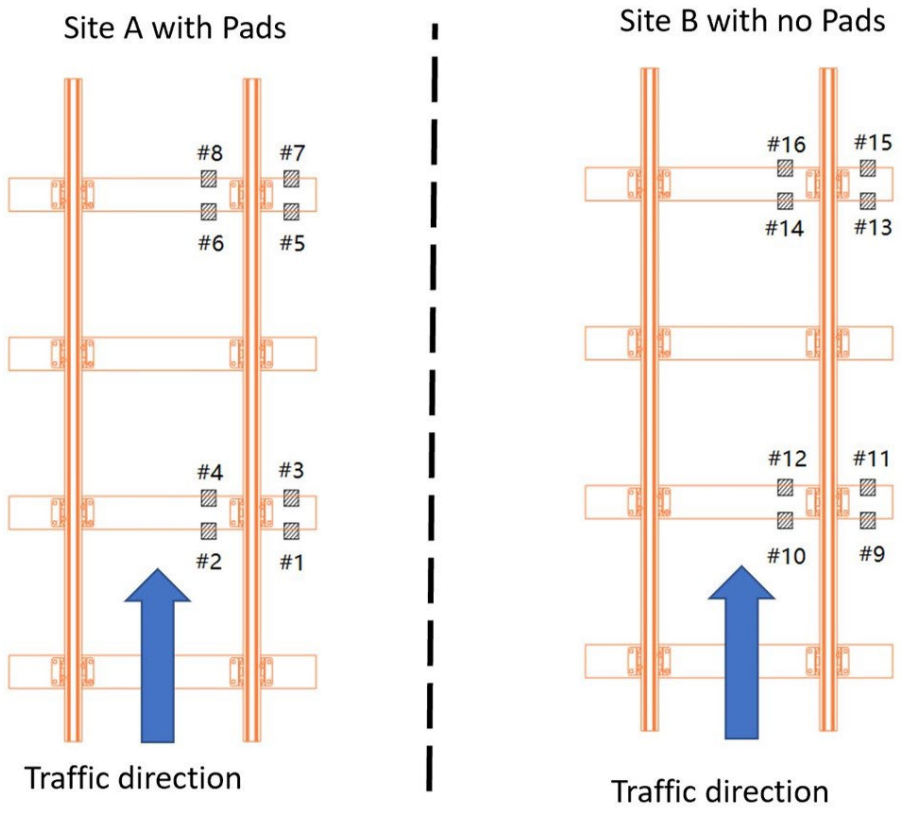


Figure 6. Diagram illustrating each SmartRock's installation location and site

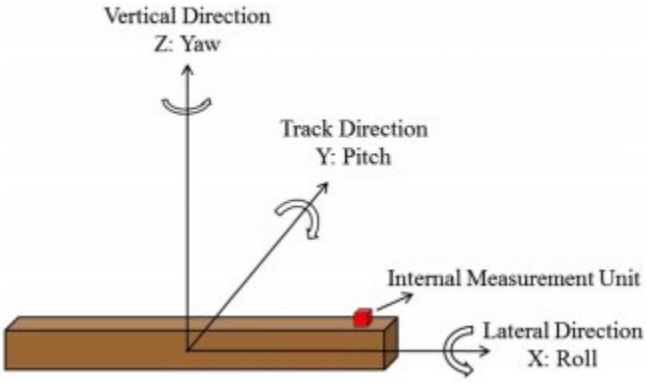


Figure 7. Global track coordinate basis

CHAPTER 3

Data

Data were collected starting on February 1, 2020 and concluded on March 9, 2020 after the SmartRock internal batteries ran dry. The first accurate and usable data were recorded on February 24, 2020. Throughout the SmartRock measurement period following this date, most SmartRocks measurements included a mix of a good and bad datasets, while some failed to measure any good data. Since all data recorded prior to February 24th were determined to be unusable by the research team tasked with installing the SmartRocks, only data recorded on or after this date were investigated.

PRELIMINARY DATA EXPLORATION

Figure 8 illustrates an example of an unusable raw dataset, showing the raw z-acceleration channel (in gravity units g) against the integer packet number channel for SmartRock 1 recorded on February 28, 2020. While these data do not appear to be intrinsically incorrect, since the z-acceleration averages a values of −1g, which is to be expected if the SmartRock were installed with a local coordinate system similar to the defined global coordinate system, the time series does not appear to have captured a train passing due to the very small variations in acceleration. Figure 9 illustrates an example of a usable raw dataset, which was recorded for the same SmartRock the day prior. This SmartRock had much larger variations in acceleration during a portion of the recording period, indicating a train passing over the track above. The rest of the time series data surrounding the train passing are noise during which the SmartRock was still recording and required removal, which is discussed later in this chapter.

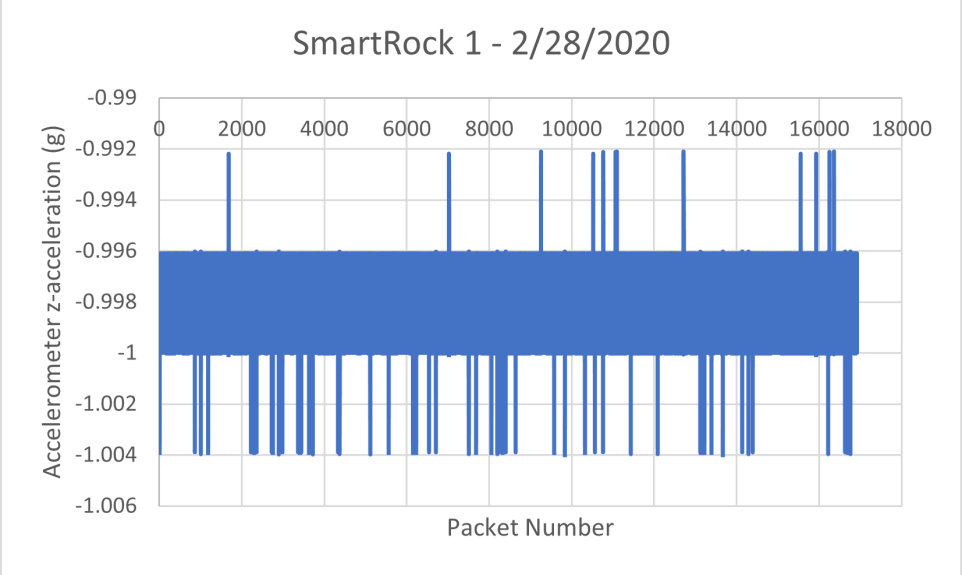


Figure 8. Raw z-acceleration in local coordinates (no passing).

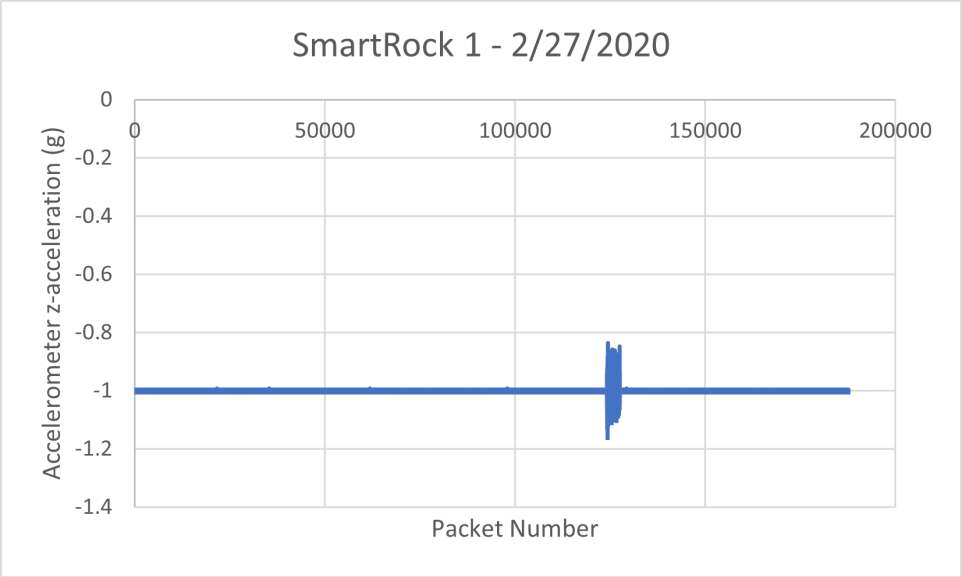


Figure 9. Raw z-acceleration in local coordinates (passing).

All the raw SmartRock acceleration data received were analyzed to determine if they appeared to have recorded a train passing or not before data pre-processing began. Table 3 tabulates all received data as being usable (green shaded cells) or unusable (red and yellow shaded cells). Most of the SmartRocks recording dates denoted in red seemed to only measure noise with their accelerometer, like what was introduced in Figure 8. Some of the SmartRocks did not work properly and thus no data were collected. SmartRocks 5 and 10 only recorded one train passing, which was not used in most analyses since their measurements were determined to be outliers under various metrics. Without multiple datasets from these SmartRocks, differences in measurements were not well understood and thus these datasets were not used in the core analyses presented in this report.

The SmartRock quaternions were also analyzed to investigate their general trends. Unit quaternions are four-value vectors that represent rotation on a real basis, mathematically defined as the quotient between two vectors in three-dimensional space (Hamilton, 1843). The x, y, and z components (the 2nd, 3rd, and 4th quaternion elements) of the quaternion represent the vector about which a rotation will occur while the 1st element represents the magnitude of the rotation (Hamilton, 1853). Thus, a time series of quaternions represents a changing coordinate basis in time which, in this case, is the local coordinate basis of the SmartRock of interest. Figure 10 and Figure 11 show time series of quaternion elements 1 and 2 of the same SmartRock dataset for which acceleration data are shown in Figure 9. From Figure 11 it is easy to see that instantaneous changes in the SmartRock's orientation are very minor until the train passing occurs, after which they increase due to dynamic wheel impacts. Similarly, Figure 10 shows a large change during the train passing, signifying the axis around which the rotation is occurring has changed.

Table 3. Available SmartRock data; green cells are good datasets.

Date	Site A									Site B								
	1	2	3	4	5	6	7	8	Strain guage	9	10	11	12	13	14	15	16	Strain guage
2020-02-21																		
2020-02-22																		
2020-02-23																		
2020-02-24	2	2				2	2	2	2			2		2	2			
2020-02-25																		
2020-02-27	1	1				1	1	1				2		2	2			1
2020-02-28																		
2020-02-29		1				1	1	1			1	1		1	1			1
2020-03-01	1	1			1	1	1	1				1		1	1			2
2020-03-02		1					1	1				2		2	2			
2020-03-03																		
2020-03-04														1	1			x
2020-03-05																		
2020-03-06																		
2020-03-07																		
2020-03-08																		
2020-03-09		1			1	1	1	1				2		2				x
The total	4	7			0	6	7	7	2		0	10	0	11	9			4

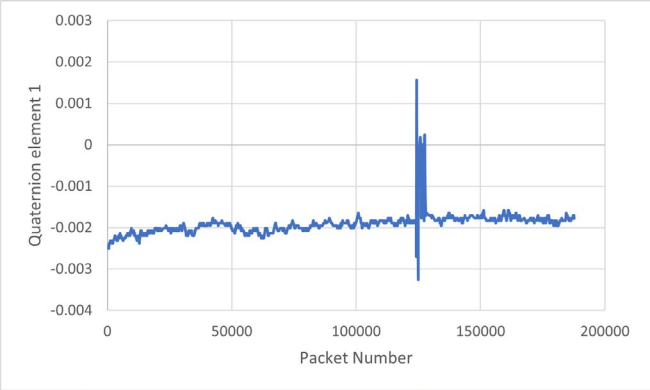


Figure 11. Quaternion Element 1.

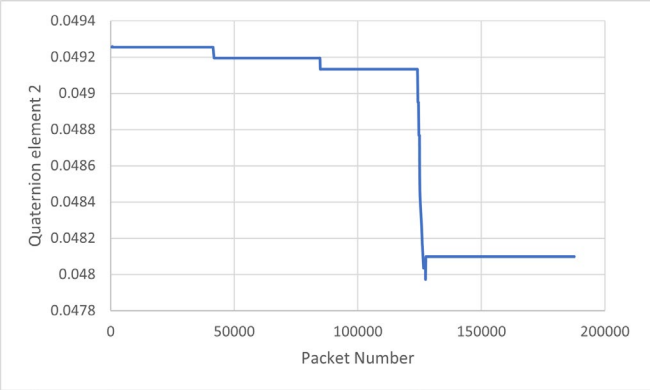


Figure 10. Quaternion Element 2.

The quaternion time series themselves were not used in analysis but were integral in changing the basis of the SmartRock acceleration data. Euler angles calculated from the quaternions describe the SmartRock's orientation with respect to the global track coordinate basis at every time step. From these angles, rotational velocity and acceleration can be calculated to analyze the SmartRock's vibration around track coordinates.

In addition to the SmartRock data, Table 3 introduces strain gauge data available as part of this dataset. Strain gauge measurement systems were installed at both testing sites as Figure 12 shows to measure how rail strain varied for different passing trains. Four strain gauge rosettes were attached to the field side of the rail above a tie between the two ties under which SmartRocks were installed. Unfortunately, much like the SmartRock data, the strain gauge data did not always measure properly and are thus incomplete. They do, however, provide some further insight into the trains that passed over the two track sites; in particular, by allowing estimates of the train speeds to be made. Figure 13 illustrates an example of strain measurements, in this particular case measured on 2/27/2020 at Site B. There were no recording dates in which the strain gauge data were measured properly at both track sites in this dataset, so cross-comparisons were not

possible. The plot clearly shows the time at which a train passed on this date, with strain values increasing dramatically.

Figure 12. Four strain gauge channels attached to the field side of the rail.

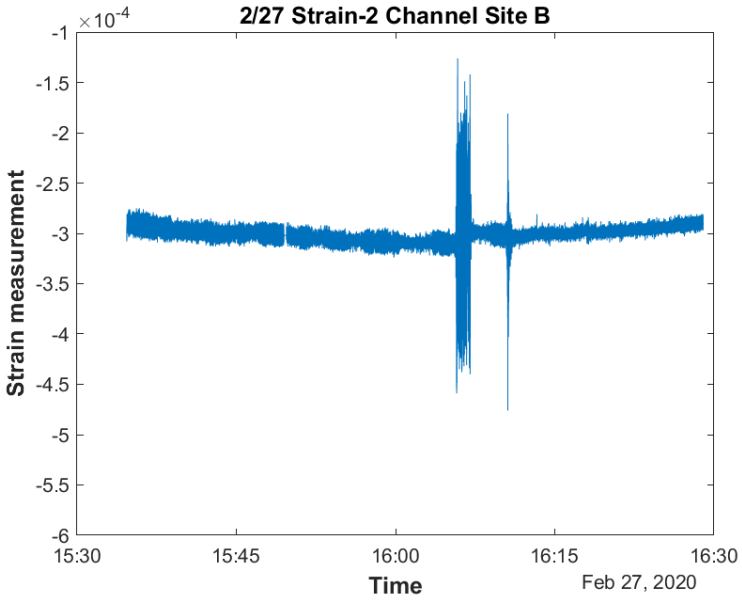


Figure 13. Plot of strain data, measured by channel two, at track Site B.

On a macro-scale, the strain gauge data were not useful due to the limited quantity. They did, however, allow estimates of train speeds to be made for a selection of dates on which the data were properly measured, by using a peak analysis algorithm and correlating it with assumed load impulses. If a distance between the load impulses can safely be assumed, and that distance is correlated with the time step seen between strain peaks in Figure 14, a train speed can be estimated. Estimated train speeds from the available strain gauge data are shown in Table 4. Even though the estimated train speeds do not encompass the entire SmartRock dataset, since the strain gauge data were incomplete, they were used to compare certain aspects of the SmartRock dataset. Additional information from this table, such as the mean value of strain peaks, was used to investigate how strain gauge and SmartRock measurement correlated.

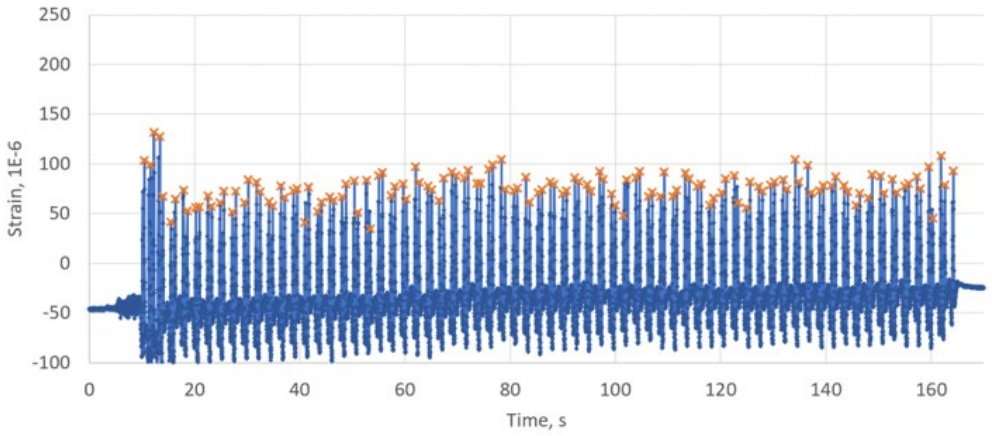


Figure 14. Strain peaks calculated for an example strain gauge series.

Table 4. Train speeds as estimated from the available strain gauge data.

Loading and environmental condition

Dates	Strain Peak Value				Train Speed (mph)	Temp(F°)		Precipitation (inch)
	Mean	Std	Max	Min		Max	Min	
02-24 (1)	277.81	18.79	305.44	236.30	19.62	55	41	0
02-24 (2)	240.43	38.51	328.55	174.51	20.98			
02-25	163.00	15.43	208.79	146.90	10.51	43	32	0.1
02-27	228.56	55.11	338.61	98.20	21.21	61	25	0
02-29	198.19	27.31	249.53	180.06	16.53	75	30	0
03-01	252.32	39.89	406.70	230.87	23.02	68	45	0

DATA PRE-PROCESSING

After determining which SmartRock datasets would be usable in further analysis and gathering all secondary sources of data recorded as part of this project, data pre-processing was performed to clean the data and transform its coordinate basis to the track coordinate basis to prepare it for further comparative analysis. Data preprocessing was performed following the steps shown below:

1. Read in acceleration and quaternion time series for the given dataset.
2. Transforms quaternion time series using the first orientation measurement of the SmartRock.

3. Interpolate quaternion and acceleration time series such that a measurement is available for every integer packet number.
4. Calculate a rotation matrix for each time step of data from the quaternion vector at that step.
5. Transform the corresponding acceleration vector into global track coordinates, per Figure 7, using the rotation matrix.
6. Extract train passings from the time series by cutting off acceleration noise and finding the proper indices that enclose the train passing.

SmartRock data were read into a MATLAB code written to perform for all preprocessing steps. A "for loop" was used to analyze each SmartRock dataset individually. After the acceleration and quaternion time series were read in, the quaternion series underwent a basis transformation such that it reconciled the first quaternion measured for that SmartRock. The first quaternion measured for a SmartRock, which for this dataset was the first measurement made on February 24, details the initial orientation of the SmartRock. Multiplying all subsequent quaternion measurements for that SmartRock by the initial measurement changes the coordinate basis to the local coordinate system of the SmartRock, which varies between SmartRocks based on how they were installed. This transformation is done per the equations below.

$$Q'(t) = Q(t) * Q(0)$$
$$[q_1'(t)q_2'(t)q_3'(t)q_4'(t)] = [q_1(t)q_2(t)q_3(t)q_4(t)] * [q_1(0)q_2(0)q_3(0)q_4(0)]$$

Where

$Q(t)$ = quaternion vector at time t
$Q'(t)$ = quaternion vector at time t after transformation
$q_i(t)$ = quaternion element i at time t

After the quaternion time series underwent their basis transformation, the time series in the current dataset were linearly interpolated. This includes all dimensions of both the acceleration and quaternion series. Linear interpolation is performed per the equation below.

$$y_{i+1} = y_i + (x_{i+1} - x_i)\frac{y_{i+2} - y_i}{x_{i+2} - x_i}$$

Where

y_i = dependent time series at index i
x_i = independent time series (packet numbers) at index i

Quaternion vectors were measured far less often than acceleration vectors, since the orientation of the SmartRock changes at a much slower rate. Table 5 and Table 6 show snapshots of .csv raw data files with datetime signatures, integer packet numbers, and measurements for each value of the acceleration and quaternion time series vectors, respectively. From initial inspection, a few things became clear about these time series. As mentioned before, the recorded acceleration and quaternion measurements do not line up with the recorded stream of packet numbers. Furthermore, acceleration values were measured inconsistently, with jumps of several packet numbers sometimes occurring.

It initially appears that the SmartRock acceleration measurement sample rate was two packets, but Table 5 refutes this. The sample rate starts out as two packets in this measurement sample, but a measurement is made for every packet from 42 − 50. From this, it was assumed that the sample rate for the accelerometer was one packet. This assumption was made before learning the design measurement sample rate for the accelerometer, which was 50 Hz. The sample rate in seconds that corresponds to one packet can be determined from the datetime stamps that accompany each instantaneous measurement. From Table 5, the

time step associated with an integer step of packets is not uniform but is very small (on the scale of several milliseconds). Initially, it does not appear that the measurements in this table support the known accelerometer sampling frequency of 50 Hz because of this.

Table 5. Snapshot of an acceleration .csv raw data file.

LoggedData_test1_20_02_27_14_30_Accel.csv				
Month:Day:Hr:Min:Sec:MSec	Packet number	Accelerometer X (g)	Accelerometer Y (g)	Accelerometer Z (g)
2:27:14:30:0:660	1	0.00390625	-0.00390625	-1
2:27:14:30:0:667	3	0.00390625	-0.00390625	-1
2:27:14:30:0:710	5	0.00390625	-0.00390625	-1
2:27:14:30:0:719	7	0.00390625	-0.0078125	-1
2:27:14:30:0:740	9	0.00390625	-0.0078125	-1
2:27:14:30:0:760	11	0.00390625	-0.00390625	-1.003906
2:27:14:30:0:776	13	0.00390625	-0.00390625	-1
2:27:14:30:0:797	15	0.00390625	-0.00390625	-1
2:27:14:30:0:817	17	0.00390625	-0.0078125	-1
2:27:14:30:0:834	19	0.00390625	-0.0078125	-1
2:27:14:30:0:852	21	0.00390625	-0.0078125	-1.003906
2:27:14:30:0:871	23	0.00390625	-0.00390625	-1.003906
2:27:14:30:0:891	25	0.00390625	-0.0078125	-1.003906
2:27:14:30:0:944	27	0.00390625	-0.00390625	-1
2:27:14:30:0:982	29	0.00390625	-0.0078125	-1
2:27:14:30:1:39	31	0.00390625	-0.0078125	-1
2:27:14:30:1:74	33	0.00390625	-0.00390625	-1
2:27:14:30:1:93	34	0.00390625	-0.00390625	-1
2:27:14:30:1:113	35	0.00390625	-0.0078125	-1
2:27:14:30:1:244	42	0.00390625	-0.0078125	-1
2:27:14:30:1:263	43	0.00390625	-0.0078125	-1
2:27:14:30:1:282	44	0.00390625	-0.00390625	-1
2:27:14:30:1:300	45	0.00390625	-0.00390625	-1
2:27:14:30:1:319	46	0.00390625	-0.0078125	-1.003906
2:27:14:30:1:338	47	0.00390625	-0.0078125	-1
2:27:14:30:1:357	48	0.00390625	-0.0078125	-1
2:27:14:30:1:375	49	0.00390625	-0.0078125	-1
2:27:14:30:1:394	50	0.00390625	-0.00390625	-1.003906

Table 6. Snapshot of a quaternion .csv raw data file.

LoggedData_test1_20_02_27_14_30_Quaternion.csv					
Month:Day:Hr:Min:Sec:MSec	Packet number	Element 1	Element 2	Element 3	Element 4
2:27:14:30:8:935	449	-0.002441413	0.0492555	0.998782	-0.001586918
2:27:14:30:9:176	462	-0.002441413	0.0492555	0.9987819	-0.001647953
2:27:14:30:10:378	526	-0.0025026	0.04925849	0.9987816	-0.001648054
2:27:14:30:11:820	603	-0.0025026	0.04925849	0.9987815	-0.001709093
2:27:14:30:13:506	693	-0.002441561	0.0492585	0.9987817	-0.001648054
2:27:14:30:14:950	770	-0.002380378	0.04925551	0.998782	-0.001647954
2:27:14:30:21:232	1105	-0.002380378	0.04925551	0.998782	-0.001647954
2:27:14:30:23:163	1207	-0.002319343	0.04925551	0.9987822	-0.001647954
2:27:14:30:23:406	1220	-0.002319343	0.04925551	0.9987822	-0.001647954
2:27:14:30:26:519	1383	-0.002319342	0.04925551	0.998782	-0.001708989
2:27:14:30:28:733	1501	-0.002319342	0.04925551	0.998782	-0.001708989
2:27:14:30:30:908	1617	-0.002319342	0.04925551	0.998782	-0.001708989
2:27:14:30:35:32	1837	-0.002380377	0.04925549	0.9987818	-0.001770024
2:27:14:30:35:276	1850	-0.002380377	0.04925549	0.9987818	-0.001770024
2:27:14:30:35:520	1863	-0.002380377	0.04925549	0.9987818	-0.001770024
2:27:14:30:35:764	1876	-0.002380377	0.04925549	0.9987818	-0.001770024
2:27:14:30:42:307	2222	-0.002380377	0.0492555	0.9987819	-0.001708989
2:27:14:30:42:776	2247	-0.002380377	0.04925549	0.9987818	-0.001770024
2:27:14:30:46:189	2428	-0.002319342	0.04925551	0.998782	-0.001708989
2:27:14:30:46:434	2441	-0.002319342	0.04925551	0.998782	-0.001708989
2:27:14:30:46:677	2454	-0.002319342	0.04925551	0.998782	-0.001708989
2:27:14:30:54:177	2853	-0.002258307	0.04925551	0.9987821	-0.001770025
2:27:14:30:54:421	2866	-0.002258307	0.04925551	0.9987821	-0.001770025

Linear interpolation of SmartRock measurements was determined to be a reasonable method of data pre-processing to match the orientation and acceleration time series. The acceleration time series infrequently had packet number step sizes from one instantaneous measurement to the next larger than a few packets, meaning most measurements expected for a sampling frequency of 50 Hz were present. Since the time step associated with a packet is so small, few approximations were being made by linearly interpolating the acceleration data.

Procedurally, first a matrix of packet numbers from the minimum to the maximum packet value recorded by the accelerometer was created. From this, an average timestep in seconds per packet number can be calculated using the equation below.

$$t_{step} = \frac{p(max) - p(1)}{t(max) - t(1)}$$

Where

t_{step} = average time step
$p(x)$ = packet number time series
$t(x)$ = datetime series
x = index of time series

The difference between the total seconds elapsed calculated from the first datetime measurement and the last makes up the denominator of the above equation. The following equation details how the total seconds elapsed since the start of the day was calculated. Using the calculated average time step *tstep*, an additional matrix was created of time in seconds starting at zero for each interpolated packet number value. This made plotting the acceleration data against a time component, rather than the packet number series, in the secondary EDA and future analysis easier, since a series detailing time was now available as a double.

$$t(x) = 60^2 * Hr(x) + 60 * Min(x) + Sec(x) + \frac{1}{1000} Msec(x)$$

Where

$Hr(x)$ = hour of datetime stamp at index x
$Min(x)$ = minute of datetime stamp at index x
$Sec(x)$ = second of datetime stamp at index x
$Msec(x)$ = millisecond of datetime stamp at index x

Next, the acceleration and quaternion time series were interpolated to match the interpolated packet number matrix. For the acceleration time series, this process was straightforward, since raw data are always available for the first and last value in the time series due to the decision to use the raw packet number matrix from the accelerometer recordings to create the interpolated packet number matrix. Linear interpolation was used due to the large count of data points in the SmartRock time series and the accompanying small time step (which is around 19 msec for most of the SmartRock data). This supports the design SmartRock sampling frequency of 50 Hz provided by the researchers tasked with installing this set of SmartRocks.

The quaternion time series were also linearly interpolated. Since the packet numbers of the first and last quaternion measurements were often different than those of the acceleration measurements, as Table 5 illustrates, a simple linear interpolation alone cannot be used. In addition to linear interpolation, missing quaternion values from the minimum packet number recorded in the acceleration time series to the minimum packet number recorded in the quaternion time series were filled in using the first recorded quaternion in the time series. For the example spreadsheets introduced in Table 5 and Table 6, this means that the quaternion vectors for packets $1 - 448$ were the same as the quaternion vector measure for packet 449. Similarly, missing quaternion vectors for packets past the maximum recorded in the quaternion time series to the maximum packet recorded in the acceleration time series were equivalent to the last quaternion vector recorded in the time series. This step is moreso a formality to equalize the length of all time series such that subsequent time series analysis techniques can be utilized, for which a constant time step is required. Train passings are a small subset of the full time series, so the quaternion series portion that make up the ends of the quaternion time series are inconsequential.

Following the time series interpolations, the acceleration data underwent a basis transformation such that it was in the global track coordinate basis seen in Figure 7 with x acceleration along the tie, y acceleration along the track, and z acceleration vertical, perpendicular to the track plane. Following the transformation of the quaternion time series to the SmartRock local coordinate system defined by the orientation of the SmartRock when it was initially installed, rotating the SmartRock acceleration data to the global track coordinate basis is as simple as applying a rotation matrix calculated from the quaternion to the acceleration data in local coordinates per the equation below, analogous to a linear transformation in linear algebra (Gentle, 2007).

$$R = \begin{bmatrix} 1-2(q_3^2+q_4^2) & 2(q_2q_3-q_1q_4) & 2(q_2q_4+q_1q_3) \\ 2(q_2q_3+q_1q_4) & 1-2(q_2^2+q_4^2) & 2(q_3q_4-q_1q_2) \\ 2(q_2q_4-q_1q_3) & 2(q_3q_4+q_1q_2) & 1-2(q_2^2+q_3^2) \end{bmatrix}$$

$$A_{global} = RA_{local}$$

Where

R = rotation matrix
A = acceleration time series

SmartRocks were calibrated such that the global track coordinate basis is the origin and the orientation of the SmartRock is measured with respect to it, as shown in Figure 7. Assuming the SmartRocks were calibrated accurately, the acceleration data in the track basis (A_{global}) can be calculated from the above equations. For each time step, a rotation matrix *R(t)* is calculated from the transformed quaternion *Q(t),* which is then applied to the 3-dimensional acceleration vector *A(t)*, putting it in the track coordinate basis. This is done for all time steps per the interpolated packet number matrix, thus ensuring the entire acceleration time series is now in the global coordinate basis ready for further evaluation.

To analyze changes in the SmartRocks orientation over time, Euler angles were calculated from the quaternion vectors that define the SmartRocks orientation with respect to the track coordinate basis, equivalently to the quaternions, but in an easier-to-analyze form. Same as the acceleration time series, an Euler angle time series was calculated for each dataset that spans the interpolated packet number matrix. The equations below define how to calculate the three Euler angles around the global track axes, shown in Figure 7, from the quaternion vectors. Like the quaternion vectors displayed in Table 6, the Euler angles vary very little when there is no train passing overtop the SmartRock but vary significantly when there is, signifying more volatile SmartRock vibration. The acceleration data in the track coordinate basis and the Euler angles detailing the SmartRock local coordinate system with respect to the track coordinate system are the two defining types of data featured in the analysis presented in this report.

$$\phi = arctan2(2(q_3q_4 - q_1q_2), 2(q_1^2 + q_4^2) - 1)$$

$$\theta = arctan\left(\frac{2(q_2q_4 + q_1q_3)}{\sqrt{1 - 4(q_2q_4 + q_1q_3)^2}}\right)$$

$$\psi = arctan2(2(q_2q_3 - q_1q_4), 2(q_1^2 + q_2^2) - 1)$$

Where

ϕ = "Roll" around the lateral track axis
θ = "Pitch" around the longitudinal track axis
ψ = "Yaw" around the longitudinal track axis

The last pre-processing step was to extract the train passings found in each SmartRock dataset. Most time series only had one train passing recorded within them, but some had two per Table 3. While more sophisticated methods were considered to extract the train passing, a simple index passing algorithm was created to extract the train passing. For time series with a single train passing, the *x* acceleration time series *x(i)*, where *i* is the index of the time series, was differenced per the equation below and 10 data points were removed from each end to eliminate any misbehaving data such as large acceleration spikes that were not discernibly part of a train passing.

$$x'(i) = x(i) - x(i-1)$$

Afterwards, the index (packet number) of all data values in the differenced time series with a larger magnitude than three standard deviations above the mean were documented. Due to the large amount of noise in the SmartRock acceleration time series, the mean of the *x* acceleration series was at or very close to zero with a relatively small standard deviation. Thus, saved data indices were always data points that were a part of acceleration spikes associated with the train passing. The minimum and maximum indices were saved and used to subset the acceleration and Euler angle time series of that dataset to just include the train passing. Figure 16 shows raw *z* acceleration data and Figure 15 shows the same data on the right after being transformed into the track coordinate basis, subsetted to isolate the train passing, and mean-removed, which was done for all acceleration time series.

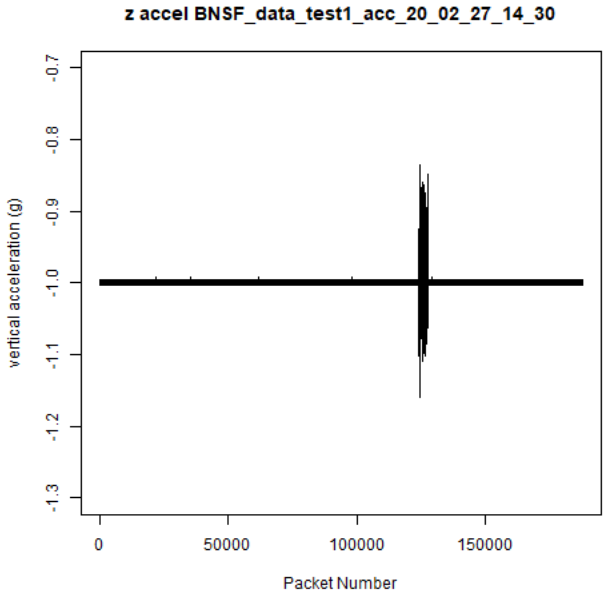


Figure 16. Full time series.

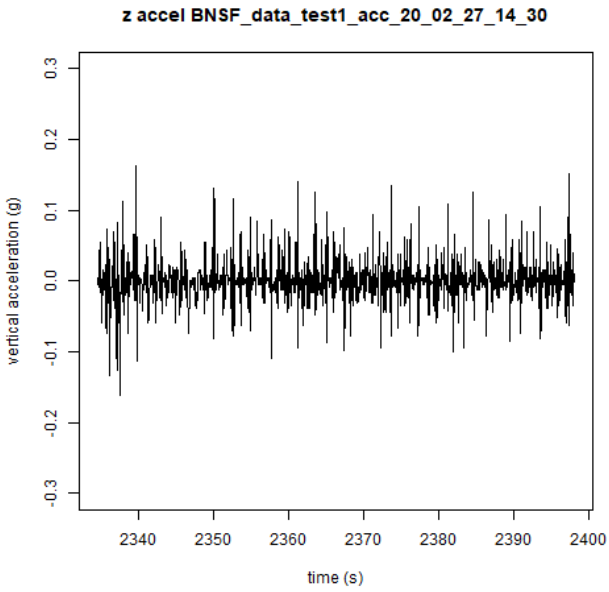


Figure 15. After pre-processing.

This procedure was slightly modified to account for datasets with two measured train passings. All steps up to and including the basis transformation were identical, after which the time series were split into two separate datasets to divide the individual train passings. Once split, train passings were extracted from the two resulting time series by calculating all data value indices that were significantly larger in magnitude than the mean and spanning the dataset using the minimum and maximum indices.

PRELIMINARY DATA EXPLORATION

An exploratory data analysis began after the acceleration data were transformed into the track coordinate basis and Euler angle time series were calculated. To verify the basis transformation, simply looking at the mean of the three acceleration components would be sufficient. If the vertical acceleration has a mean of −1g, due to the constant gravitational force acting on the SmartRock, and the lateral and longitudinal acceleration series have a mean of zero, then we can assume the SmartRock acceleration data were properly aligned into the track coordinate basis. Table 7 shows that most SmartRock datasets showcase the correct behavior, with *z* acceleration meaning very close to −1. Notably, a handful of the vertical acceleration means have a magnitude significantly larger than the other datasets, but the datasets were all recorded on March 9, 2020.

Table 7. Mean of each z acceleration time series.

Filename	z accel mean	Filename	z accel mean
test1_acc_20_02_27_14_30	-1.000	test14_mt_1_acc_20_02_27_15_37	-1.002
test1_acc_20_03_01_14_32	-1.001	test14_mt_1_acc_20_03_02_15_30	-1.000
test1_mt_1_acc_20_02_24_15_35	-1.002	test14_mt_2_acc_20_02_27_15_37	-1.001
test1_mt_2_acc_20_02_24_15_35	-1.003	test14_mt_2_acc_20_03_02_15_30	-1.000
test11_acc_20_02_24_16_00	-0.988	test2_acc_20_02_27_14_30	-0.999
test11_acc_20_02_24_16_40	-0.982	test2_acc_20_02_29_14_30	-1.000
test11_acc_20_02_29_15_30	-0.992	test2_acc_20_03_01_14_32	-1.001
test11_acc_20_03_01_15_48	-0.990	test2_acc_20_03_02_14_30	-1.001
test11_mt_1_acc_20_02_27_15_37	-0.996	test2_acc_20_03_09_15_30	-1.003
test11_mt_1_acc_20_03_02_15_30	-0.992	test2_mt_1_acc_20_02_24_15_35	-0.998
test11_mt_1_acc_20_03_09_16_29	-1.612	test2_mt_2_acc_20_02_24_15_35	-0.998
test11_mt_2_acc_20_02_27_15_37	-0.994	test6_acc_20_02_27_14_30	-0.992
test11_mt_2_acc_20_03_02_15_30	-0.992	test6_acc_20_02_29_14_30	-0.990
test11_mt_2_acc_20_03_09_16_29	-1.614	test6_acc_20_03_01_14_32	-0.985
test13_acc_20_02_24_16_00	-1.003	test6_acc_20_03_09_15_30	-0.979
test13_acc_20_02_24_16_40	-1.003	test6_mt_1_acc_20_02_24_15_35	-0.999
test13_acc_20_02_29_15_30	-1.001	test6_mt_2_acc_20_02_24_15_35	-0.999
test13_acc_20_03_01_15_48	-1.007	test7_acc_20_02_27_14_30	-1.023
test13_acc_20_03_04_15_30	-1.002	test7_acc_20_02_29_14_30	-1.014
test13_mt_1_acc_20_02_27_15_37	-1.004	test7_acc_20_03_01_14_32	-1.018
test13_mt_1_acc_20_03_02_15_30	-1.001	test7_acc_20_03_02_14_30	-1.019
test13_mt_1_acc_20_03_09_16_29	-1.323	test7_acc_20_03_09_15_30	-1.021
test13_mt_2_acc_20_02_27_15_37	-1.004	test7_mt_1_acc_20_02_24_15_35	-1.003
test13_mt_2_acc_20_03_02_15_30	-1.002	test7_mt_2_acc_20_02_24_15_35	-1.003
test13_mt_2_acc_20_03_09_16_29	-1.322	test8_acc_20_02_27_14_30	-1.004
test14_acc_20_02_24_16_00	-1.003	test8_acc_20_02_29_14_30	-1.003
test14_acc_20_02_24_16_40	-1.004	test8_acc_20_03_01_14_32	-1.003
test14_acc_20_02_29_15_30	-1.002	test8_acc_20_03_02_14_30	-1.001
test14_acc_20_03_01_15_48	-1.003	test8_acc_20_03_09_15_30	-1.005
test14_acc_20_03_04_15_30	-1.001	test8_mt_1_acc_20_02_24_15_35	-1.004
		test8_mt_2_acc_20_02_24_15_35	-1.004

Table 7 also shows all 61 core SmartRock datasets, each of which contains only a single train passing, that were laid out in Table 3. The filename column describes the SmartRock number, recording data, and recording time as the equation below details. These 61 datasets are the core of all EDAs presented in the remainder of this section and the results that follow. Unless mentioned otherwise, all data discussed are the SmartRock time series after data pre-processing was executed. Additionally, the acceleration and Euler angle time series were mean removed for the bulk of the EDA to explore time series characteristics with a nominal running mean.

$$test1\_acc\_20\_02\_27\_14\_30 = SmartRock1\_Acceleration\_Year\_Month\_Day\_Hour\_Minute$$

DATA STATIONARITY

First, it is most appropriate to discuss the SmartRock data from a macro perspective. Due to the method of data interpolation, the acceleration and Euler angle time series are more specifically discrete-time series datasets. This was not initially the case due to the inconsistent time steps for a given step in packets observed in Table 5 and Table 6. Equalizing the time step for a given SmartRock dataset makes time series analysis easier and more consistent. It is worth noting that due to the way the time series of time stamps in seconds was calculated by using the first and last recorded datetime stamps that the time steps across the 61 datasets are not necessarily equivalent. The time steps for each dataset were documented and are presented later in this section.

Since this report predominantly focuses on time series data, it is important to investigate what type of time series data is being worked with. As initial observations of Figure 15 suggest, the train passing subset of the acceleration time series is randomly distributed, or stationary, upon initial inspection. A stationary time series is a stochastic dataset whose unconditional joint probability is not dependent on time (Gagniuc, 2017). This also means that parameters such as the time series mean and variance are not time dependent. Exploration of data stationarity assumes that only the train passing itself is being explored, since variance is time dependent of the acceleration noise. To further verify acceleration time series stationarity, an autocorrelation plot was created. Autocorrelation is the correlation of a signal (in this case an acceleration time series A) with a delayed copy of itself $\bar{A}$ as a function of delay (in this case the number of time steps) (Dunn & Davis, 2005; Papoulis, 2002). The equation below describes the function used to calculated autocorrelation. The function *acf()* in R was used to perform autocorrelation function calculations (R Core Team, 2020). Numerous packages in R were used throughout the analysis (Kuhn, 2020; Milborrow, 2020; Wei & Simko, 2017; Wickham et al., 2019; Wuertz, 2020).

$$R_{AA}(\tau) = E[A_t \bar{A}_{t+\tau}]$$

Where

A_t = acceleration time series
$\bar{A}_{t+\tau}$= complex conjugation of acceleration time series

R = autococorrelation function
E = expected value operator
τ = count of delay steps

Figure 17 shows the autocorrelation function calculated and plotted in R for the vertical acceleration time series seen plotted in Figure 15. The autocorrelation function is seen on the y-axis while the number of lag steps in the time series is seen on the x-axis. An example reading of this plot is that ACF(x = 0) = 1.0, meaning that for zero lag steps, the time series is perfectly correlated. This makes sense, since the function is correlating the time series with itself without any delay. As the number of lag steps increases, a stationary time series is exceedingly unlikely to correlate with itself. Autocorrelation function values rapidly approach zero after a few lag steps and remain there for increasingly larger lag steps when calculated for stationary processes (Dunn & Davis, 2005). The autocorrelation plots show that the calculated autocorrelation function values fall below the significance limits (which values underneath can be considered statistically insignificant) after less than a dozen lag steps and remain there throughout most of the remaining lag steps, indicating that these acceleration time series can be considered stationary with some autoregressive terms present (Papoulis, 2002). While the amount of autocorrelation does vary for the domain used in this plot, after a few lag steps it was considered to be nominal. Thus, it is safe to assume that this acceleration time series is stationary. In general, the autocorrelation for any given SmartRock acceleration time series dimension was nominal, as Figure 19 and Figure 18 display with the x and y acceleration components of

the same dataset, respectively. For the purposes of this report, after reviewing the autocorrelation functions of all acceleration time series for all three components, all acceleration data were determined to be stationary.

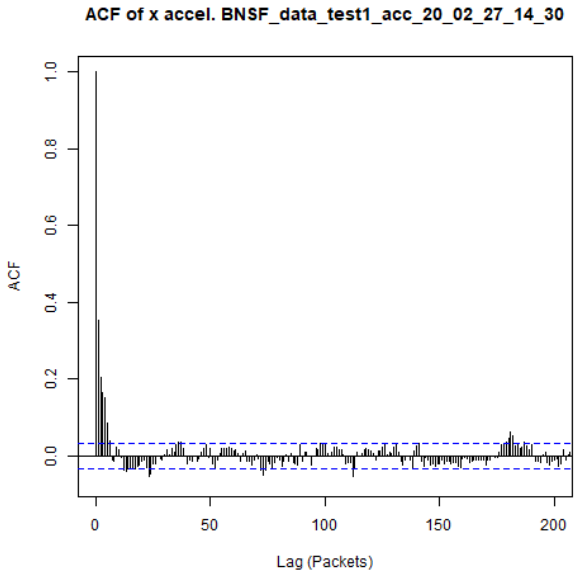


Figure 19. ACF x-component.

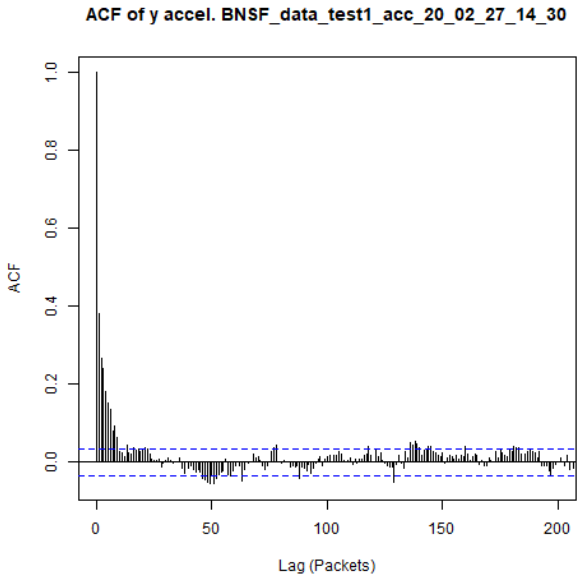


Figure 18. ACF y-component.

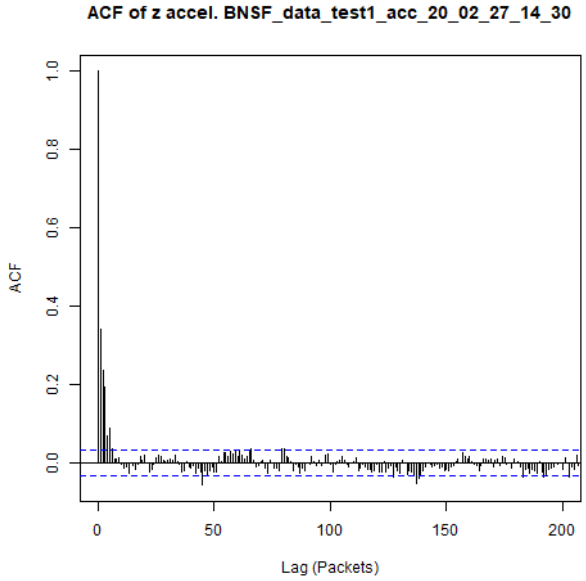


Figure 17. ACF z-component.

The Euler angle time series were also analyzed to determine if they were stationary. Figure 22, Figure 21, and Figure 20 show the Roll, Pitch, and Yaw of the train passing recorded by SmartRock 1 on February 27, 2020, respectively. From these plots, it is immediately clear that the Euler angle time series are not stationary, since there are clear trends in the plots throughout the train passing. While not stationary, the Roll and Pitch time series have a reasonably random shape. Conversely, the Yaw time series increases with time and has a clear mean trend. To verify, the data were differenced (a typical method of coercing a time

series into stationarity) and fed into the autocorrelation R function to see if the plots told a different story (R Core Team, 2020).

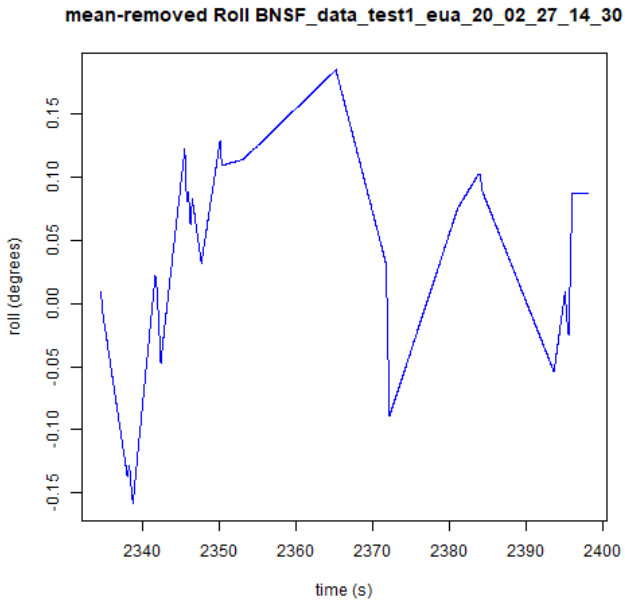


Figure 22. Roll.

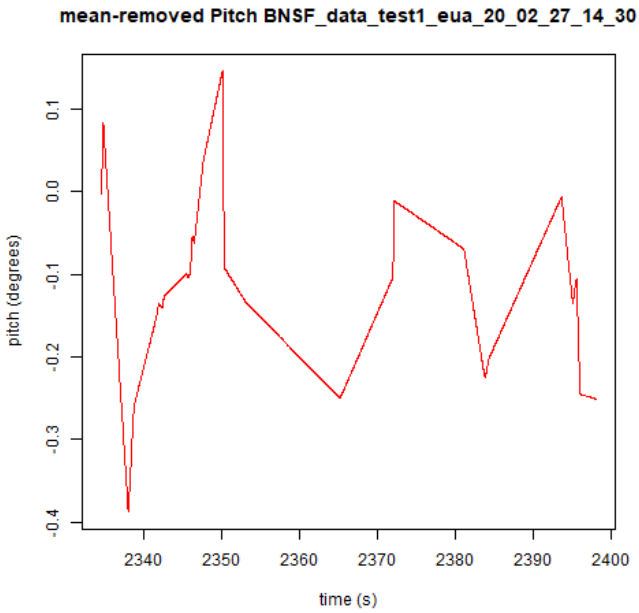


Figure 21. Pitch.

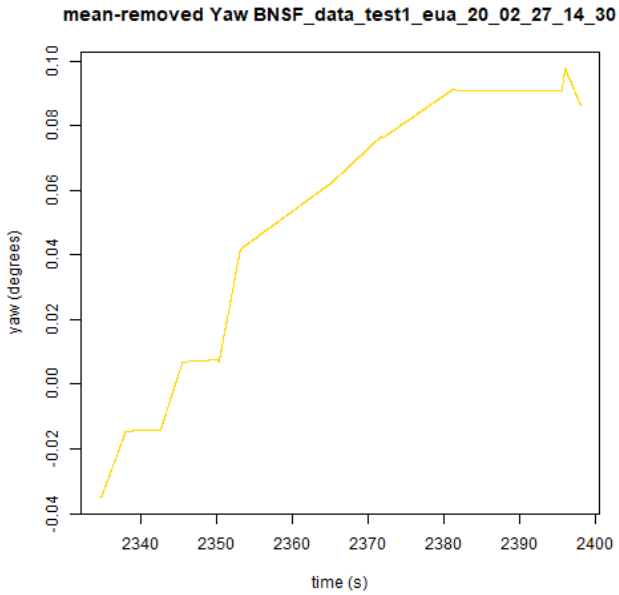


Figure 20. Yaw.

As expected, Figure 25, Figure 24, and Figure 23 show clear autocorrelation in all three differenced Euler angle time series at various lag steps. Similarly to the acceleration data, all autocorrelation functions of the differenced Euler angle time series were examined and it was determined that the data should not be considered stationary for the sake of analysis introduced in this report.

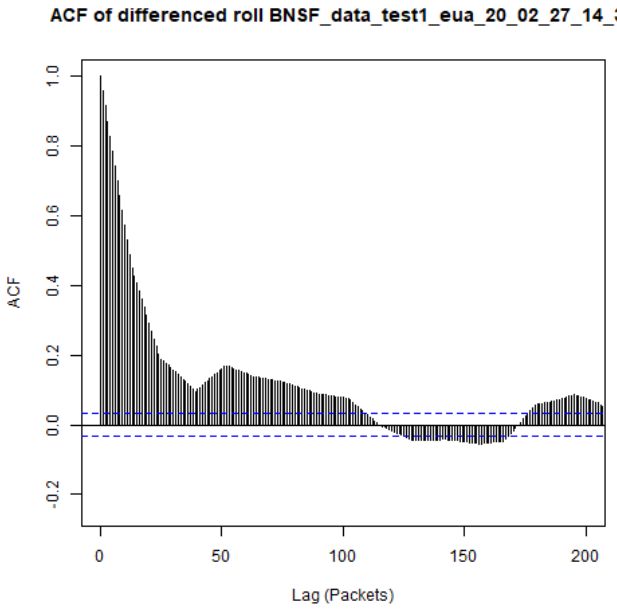


Figure 25. ACF of Roll.

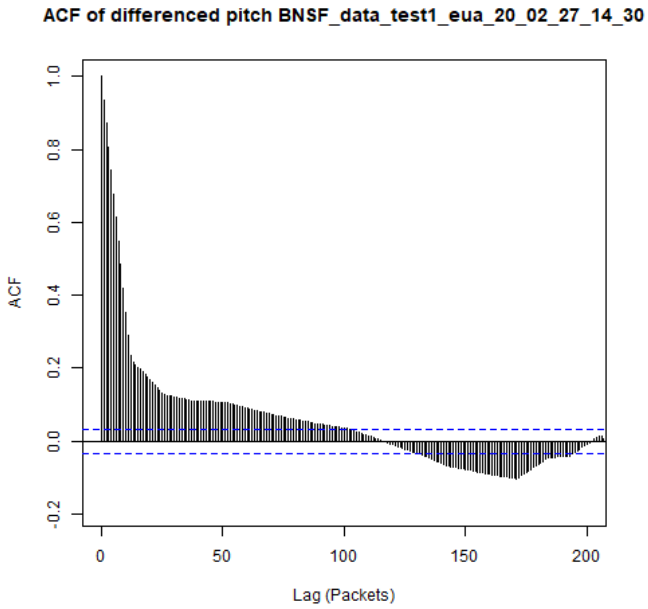


Figure 24. ACF of Pitch.

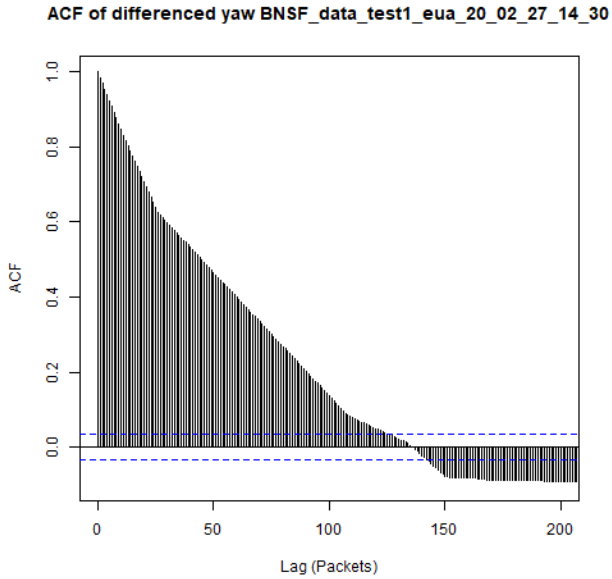


Figure 23. ACF of Yaw.

TIME SERIES DISTRIBUTIONS

With a better understanding of the general shape and behavior of the time series available in this SmartRock dataset, a more specific EDA can be performed with confidence. Part of this EDA was to examine the distribution of acceleration and Euler angle data values with respect to each individual dataset. Knowing that the acceleration data are stationary indicates that they are likely normally or near normally distributed. Numerous methods of calculating and displaying a time series distribution were used to investigate the values in each time series and how they differed between datasets, all of which are presented and explained

in this section. In general, investigating how the time series data values were distributed helped guide analysis of the SmartRock data conducted for this research report.

The first type of distribution calculated for the acceleration time series data was histograms. They were calculated for all 61 acceleration time series in all three dimensions after they were mean removed. Based on the stationarity conclusion made in the previous section, it was expected that these histograms would be distributed normally with a mean of zero. Figure 26 shows the histogram of the train passing seen in Figure 15, which acts as expected, distributed normally around zero. Notably, the histogram has a relatively narrow distribution, indicating it likely has less variance than a standard normal distribution. Histograms with equivalent bin sizes were calculated for all the acceleration time series, which acted similarly.

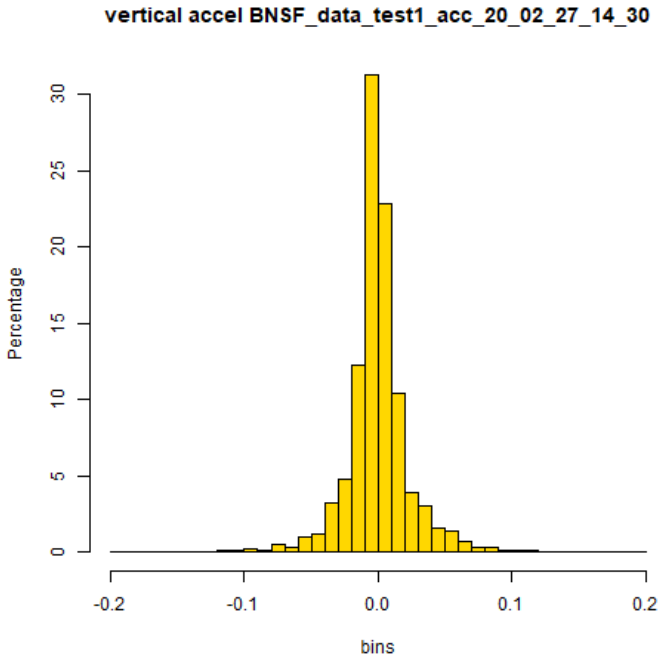


Figure 26. Histogram distribution calculated for the vertical acceleration component.

To investigate if this acceleration time series was in fact less like a standard normal distribution and more like an alternative distribution, quantile-quantile (Q-Q) distributions were calculated. Q-Q plots are a graphical method used to compare two probability distributions by plotting their quantiles against each other (Wilk & Gnanadesikan, 1968). Quantiles are values that partition a set of ordered values into q subsets, each of which has the same number of data points. The value of the quantile is the median of the set of data points in the quantile set. Q-Q plots were created for each acceleration time series to compare their distribution of values to those of a normal distribution. The *qqnorm()* function in R, which takes a sample series and compares its quantiles to those of a normal distribution, was used to create these plots (R Core Team, 2020). Figure 27 shows the quantile-quantile distribution for the vertical acceleration time series recorded by SmartRock 1 on 2/27/2020 with the standard deviation of the series divided out to normalize it. The y-axis shows the sample quantiles (aka the time series of interest), while the x-axis shows the theoretical quantiles (which for this plot is a standard normal distribution).

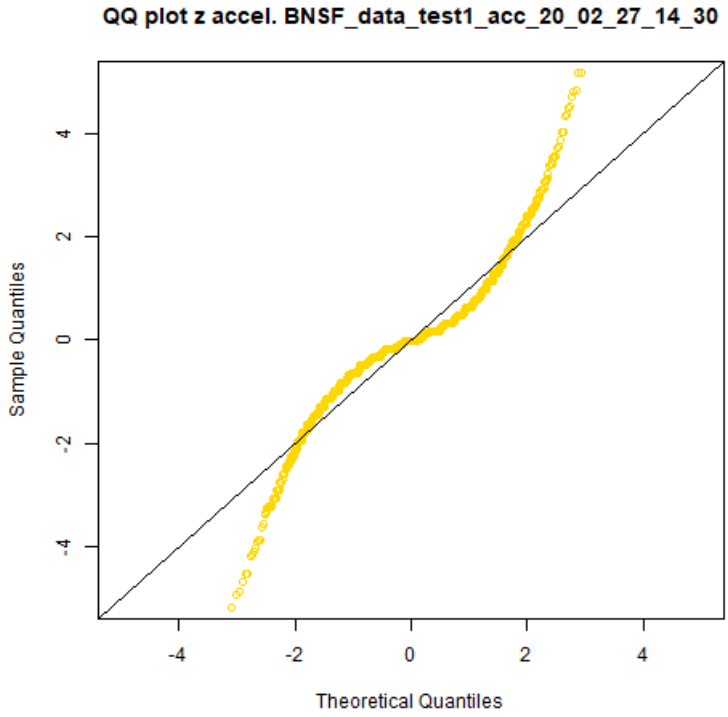


Figure 27. Q-Q distribution calculated for acceleration z-component.

If the acceleration time series is distributed normally, all quantile pairs would lie on the line $y = x$ that stretches from the bottom left to the top right of the plot. Instead, the sample quantiles were larger in magnitude than the theoretical quantiles on both ends of the distribution, indicating that the sample time series has quantiles with medians greater than three standard deviations from the mean or more while the normal distribution does not. Conversely, the sample quantiles were smaller in magnitude near the center of the distribution but align when SD = −2, 0, 2. Figure 26 and Figure 27 suggest that this acceleration time series is more likely to take the form of a continuous Laplace exponential distribution per the equation below and not a standard normal distribution (Wilson, 1923).

$$f(x|\mu, b) = \frac{1}{2b}\begin{cases} e^{-\frac{\mu - x}{b}} & if\ x < \mu \\ e^{-\frac{x-\mu}{b}} & if\ x \geq \mu \end{cases}$$

Where

$f(x|\mu, b)$ = Laplace exponential distribution
μ = location parameter
b = scale parameter

This was further investigated by creating density distributions of the acceleration time series, which disperses the mass of the empirical distribution function over a large grid and then uses the fast Fourier transform to convolve this approximation with a discretized version of the kernel and uses linear approximation to evaluate the density at the specified points per the descriptions of the *density()* R function used for the calculations (R Core Team, 2020). Essentially, this function creates a distribution of the data with minimal bin sizes. Figure 28 demonstrates this for all three acceleration time series dimensions recorded by SmartRock 1 on 2/27/2020. As expected, the vertical acceleration, and in fact all dimensions,

more accurately appear to follow a Laplace exponential distribution and not a normal distribution, with values heavy centralizing and forming a peak at the mean with long tails propagating from both ends.

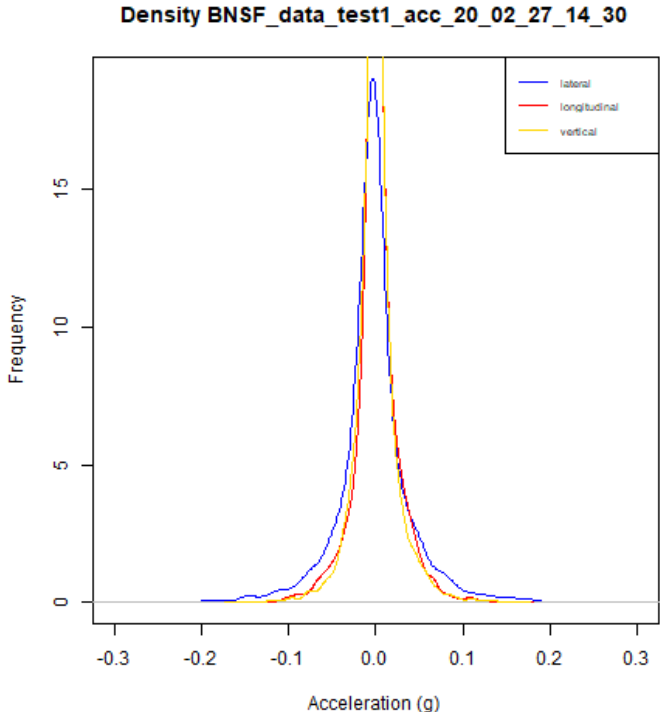


Figure 28. Probability density function for all three acceleration components.

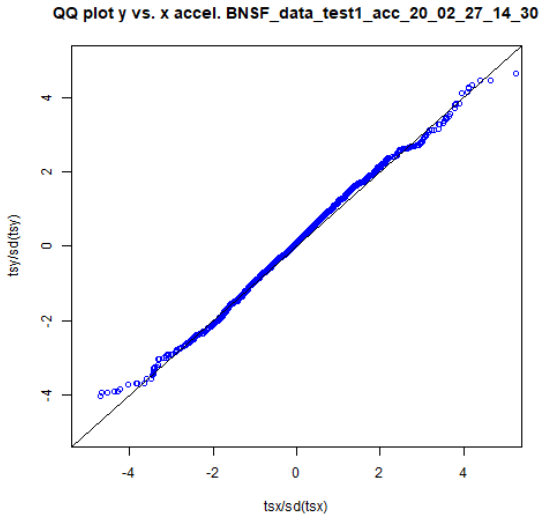


Figure 29. y vs. x.

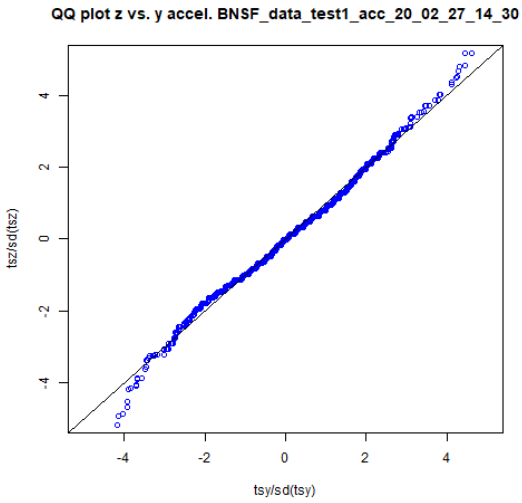


Figure 31. z vs. x.

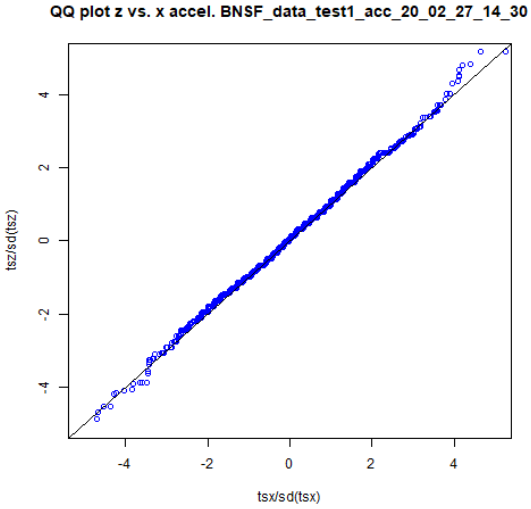


Figure 30. z vs. y.

Q-Q distributions were used again to compare the distribution of acceleration values of a dataset across dimensions after visual affirmation from Figure 28 that, based on initial inspection, they are distributed quite similarly. Q-Q plots were created to compare the *z* acceleration series to the *x* and *y* acceleration series and, additionally, the *x* and *y* acceleration series were cross-compared. The results for the train passing recorded by SmartRock 1 on 2/27/2020 are shown in Figure 31, Figure 30, and Figure 29. From these plots, it is clear that the acceleration channels are highly correlated for the bulk of the quantiles, with less correlation being seen near the tail ends of the distributions. This behavior, and the behavior displayed by the distributions introduced for the train passing recorded by SmartRock 1 on 2/27/2020 in this section, was seen broadly for all the acceleration time series in every dataset.

Before concluding this section, it is important to note that equivalent distributions were calculated and produced for the Euler angle time series as well. As expected from the results observed in Figure 22, Figure 21, and Figure 20, the Euler angle time series were not normally distributed and in fact did not take on any consistent distribution pattern. Probability density functions calculated for all the Euler angle time series recorded by SmartRock 1 on 2/27/2020 are shown in Figure 33, Figure 32, and Figure 34. From these, it is clear that the Euler angle time series do not have similar shapes to one another, nor do they have similar shapes to their acceleration density function counterparts. This is due to the irregular changes in orientation the SmartRocks appeared to undergo throughout a train passing.

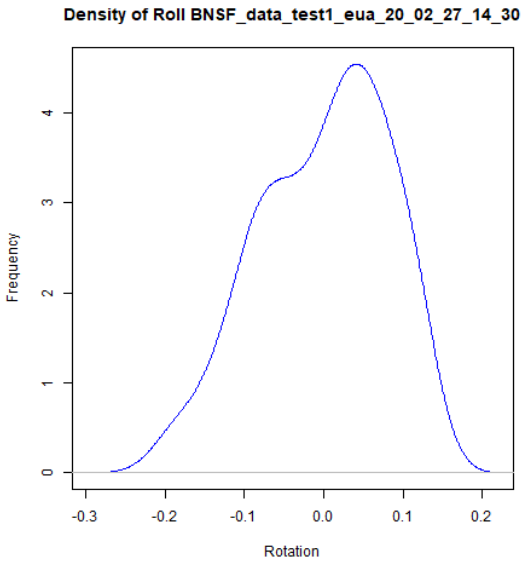


Figure 33. Density of Roll.

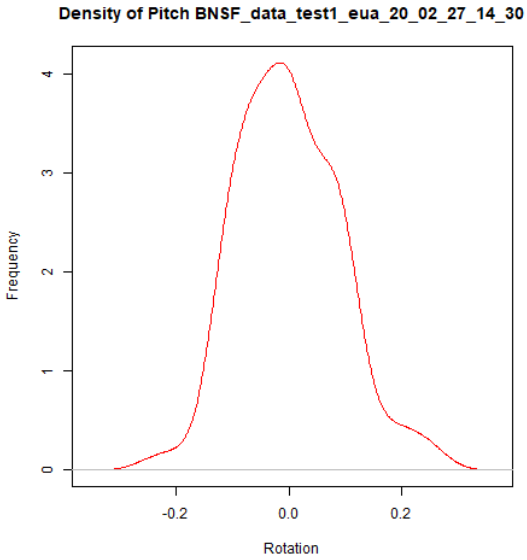


Figure 32. Density of Pitch.

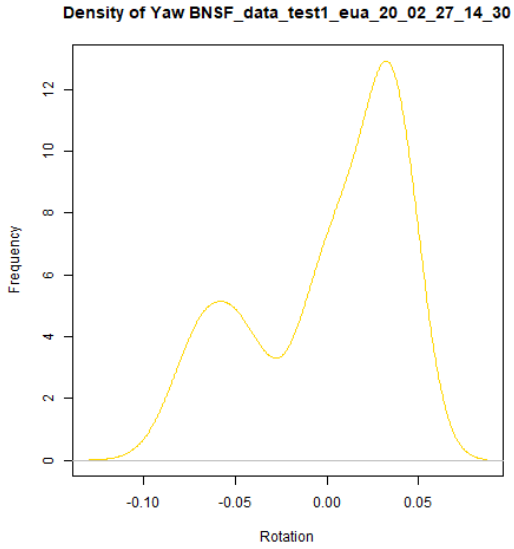


Figure 34. Density of Yaw.

CATEGORICAL DATA CLASSIFIERS

Part of this EDA was to denote the differences in the 61 available train passings. Comparing and contrasting the time series based on various categorical classifications of the datasets was a core part of the analysis of this set of SmartRock data and determining the effect of ballast condition, or other factors, on it. The list below shows all categorical classifications attached to each individual train passing to further differentiate them for future analysis. Most of the categorical classifications are self-explanatory: SmartRock number is the ID of the individual device in this dataset, SmartRock site indicates if that SmartRock is installed at the West (A) or East (B) site, and SmartRock recording date is the date when the train passing was recorded. SmartRock location indicates whether the SmartRock was gauge side or field side and whether it was in front of or behind the tie, with respect to the passing train, thus creating four different locations. Lastly, the

train passing number indicates whether the passing train was the first or second recorded on that date by that particular SmartRock. These categorical classifiers made it easier to divide up the datasets for investigative analysis and made it easier to determine which datasets were recording the same passing train.

1. SmartRock number
2. SmartRock site
3. SmartRock recording date
4. SmartRock location
5. Train passing number

DESCRIPTIVE SUMMARY STATISTICS

This section focuses primarily on determining and using the available SmartRock time series data to calculate descriptive summary statistics of the data. The time series, even after subsetting to isolate the train passing, often still includes a few thousand data points, making descriptive statistics of the datasets useful to summarize them. Many of the core analyses carried out and presented in the following chapter utilized the descriptive statistics described herein. Thus, these statistics are an important part of this report, and their calculation and reasoning are covered fully in this section.

Typical descriptive summary statistics of time series data were calculated and used extensively in most facets of the SmartRock analysis. However, some descriptive statistics were calculated but not used, since they did not carry any new information. Since the time series data were mean removed for most analysis in this report, calculating, for example, the mean of any given time series, yielded no new information. Other typical summary statistics such as standard deviation, maximum, and minimum were calculated for all three dimensions of the pair of time series and did provide new information. Time series standard deviation, particularly for the acceleration time series, helped illustrate quantitative differences in the data based on some of the categorical classifiers presented in the previous section. Additionally, calculating the maximum and minimum acceleration values in a given time series helped demonstrate the bounds of acceleration values and how the distribution of acceleration values varied between various SmartRocks and train passings.

Further research into appropriate summary statistics for this type of dataset led to the calculation of a five-number summary for each dimension of acceleration and Euler angle time series data. Thus, in addition to calculating the maximum and minimum, the lower (first) quartile, upper (third) quartile, and median were calculated for each time series. The lower quartile is the median value of the lower half (smallest values) of the time series, while the upper quartile is the median value of the upper half (largest values) of the time series. If the time series were ordered from smallest value to largest value, the first quartile statistic would be the 25th percentile, while the third quartile would be the 75th percentile. Under this description, the median is the 50th percentile. Thus, the five-number summary divides the time series into four quartiles equivalent to the quartiles discussed in the previous section. Figure 36 illustrates the five-number summary and how the data are divided into four quartiles, each containing 25% of the data in the time series.

From the five-number summary, the interquartile and total range of the time series were calculated. The interquartile range is $Q_3 - Q_1$ while the total range is *max-min*. Comparing these two calculated statistics can yield information about the presence of outliers and about the general distribution in the time series, since the interquartile range only encloses 50% of the data values. Figure 35 shows the range plotted against the quartile deviation for the three acceleration components of all 61 SmartRock datasets. In general, the two statistics are linearly related, but they are correlated far less than initially expected, with lateral acceleration series having a residual-squared value of 0.494, longitudinal acceleration series having a residual-squared value of 0.45, and vertical acceleration series having a residual-squared value of 0.32 for

a linear best fit, indicating some correlation. Residual-squared values can be understood as approximately the percentage of the data sequence variation that can be explained by the model's input, independent variable. Correlations between various variables and potential implications are discussed in the Results chapter.

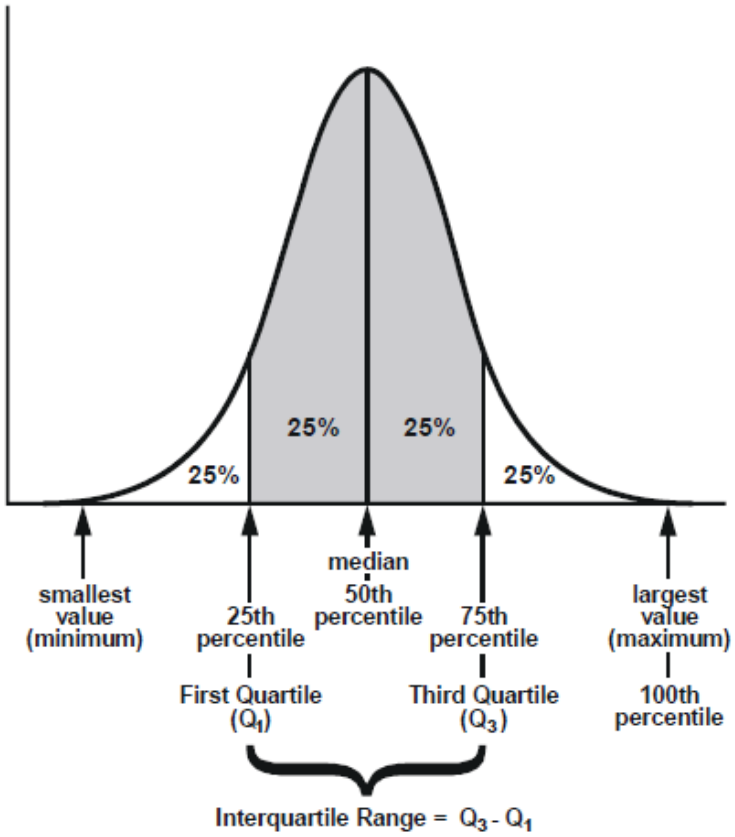


Figure 36. Diagram explaining the five-number summary.

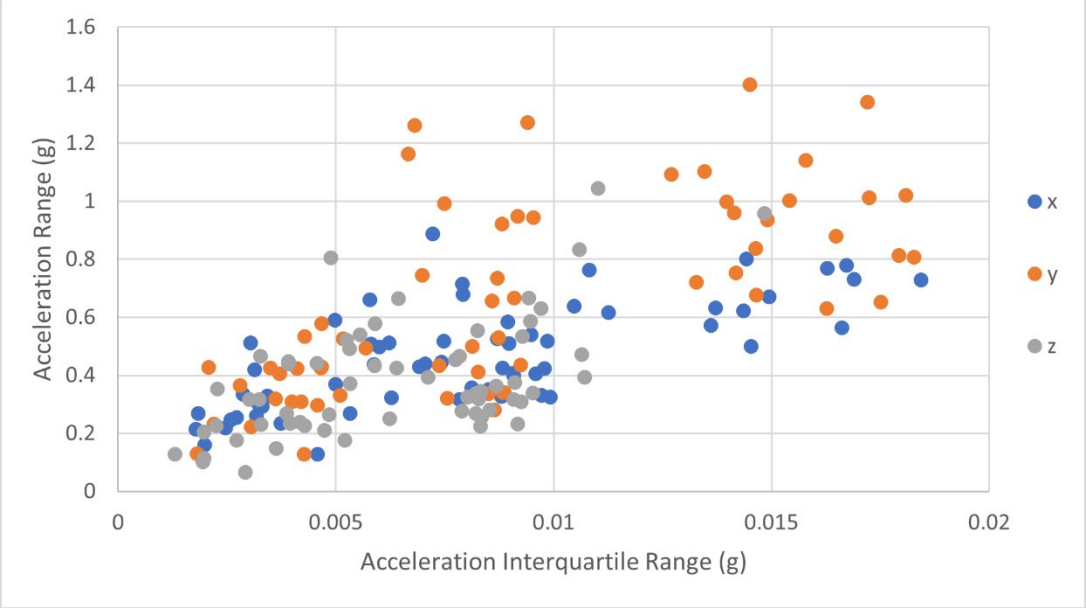


Figure 35. Acceleration max-min (range) vs. acceleration interquartile range.

There are a few other takeaways of note from this plot. On initial inspection, longitudinal acceleration time series generally have a larger range of values than the other two coordinates. Comparatively, vertical acceleration time series have smaller ranges of values and smaller interquartile ranges than the other two coordinates. The latter observation potentially indicates that SmartRocks are generally vibrating more along the lateral and longitudinal track coordinates than the vertical one. It is important to be cautious before taking this stance, however, since Figure 35 includes all available train passings, meaning different trains are included, since not all trains that passed over these sections of track over the 2½-week time frame were equivalent.

This plot also supports findings from the distributions in the previous section, that the acceleration time series most closely resemble Laplace distributions. Small interquartile ranges support that 50% of the order data are closely clustered around the mean, while the other 50% spreads much wider and sometimes produces values four or more standard deviations above or below the mean, producing a large total data range. Since the total data range relies on every data point being accurately measured, due to the fact that it only takes into account the largest positive and negative data values, the interquartile ranges were determined to be a better indicator of the variance displayed by the acceleration time series. Use of these five-number summary statistics is further explored in the Results chapter.

Other descriptive statistics were calculated in addition to the quantitative summary statistics described thus far to further quantify each of the 61 datasets in a meaningful way and to support a framework that compares the datasets in an effort to identify ballast condition from them. Calculating the number of acceleration peaks for each acceleration time series is a method identifying potential load impulses. In general, peak detection in the acceleration time series illustrates at what points in time the SmartRock was undergoing vibrations that were above average in magnitude. The indices of peaks in the acceleration time series were extracted by first differencing the time series. Afterward, all data points that had a different sign than the previous one in the time series and that had an absolute value greater than one standard deviation of the acceleration time series were determined to be peaks. The equation below shows the peak detection algorithm as written in R. Figure 39, Figure 38, and Figure 37 show the train passings recorded by SmartRock 1 on 2/27/2020 with their peaks denoted. From these plots it was clear that the indices of the peaks were calculated properly and that requiring that the peaks be one standard deviation larger in magnitude than the time series mean was reasonable. Other sensitivity levels were tested, with the resulting peaks for all acceleration time series observed, and it was determined that the sensitivity level chosen for the plots below was sufficient.

$$indices = which(sign(tsxd)! = sign(dplyr :: lag(tsxd, 1))/abs(tsx) >= sd(tsx)$$

Where

tsx = acceleration time series (lateral in this case)
tsxd = differenced acceleration time series (lateral in this case)
functions = *which(), sign(), lag(), abs(), sd()*

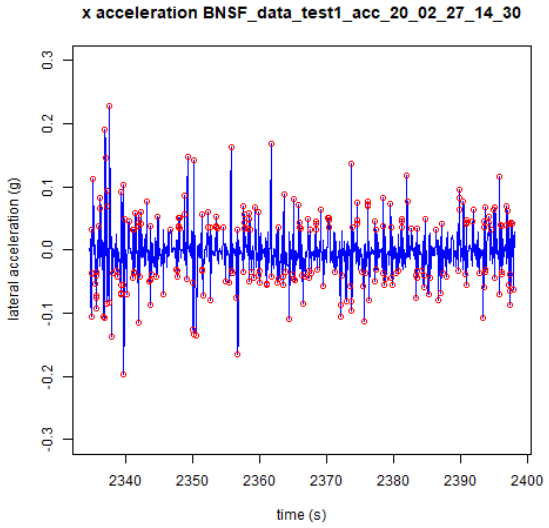


Figure 37. x peak accelerations.

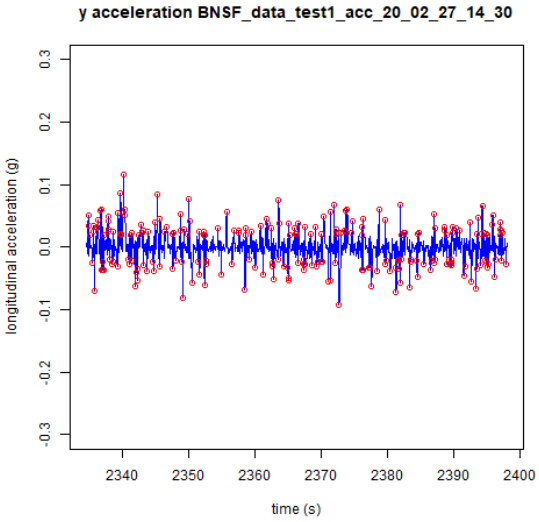


Figure 38. y peak accelerations.

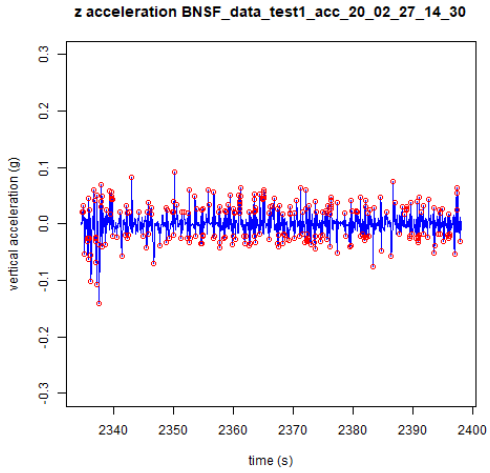


Figure 39. z peak accelerations.

The count of the acceleration peaks and their indices were saved for use in further analysis. From the above plots it appears the peaks are relatively equally spaced. Furthermore, the peaks seem to show up with similar indices for all three acceleration coordinates. Similar calculations were attempted on the Euler angle time series to determine if peaks could also be extracted from them. Unexpectedly, peaks were not meaningfully extracted due to the shape these time series tended to take. Figure 20 demonstrates why this is the case, with the Yaw of SmartRock 1 increasing linearly throughout the train passing. The Euler angle time series demonstrated various shapes throughout this dataset, so peak calculations varied greatly between them. Specifically, the Yaw of the SmartRocks throughout train passings did not demonstrate behavior that yielded any interesting observations. Suggestions by developers and experienced analyzers of SmartRock data resulted in these time series being used for little of the analysis that follows.

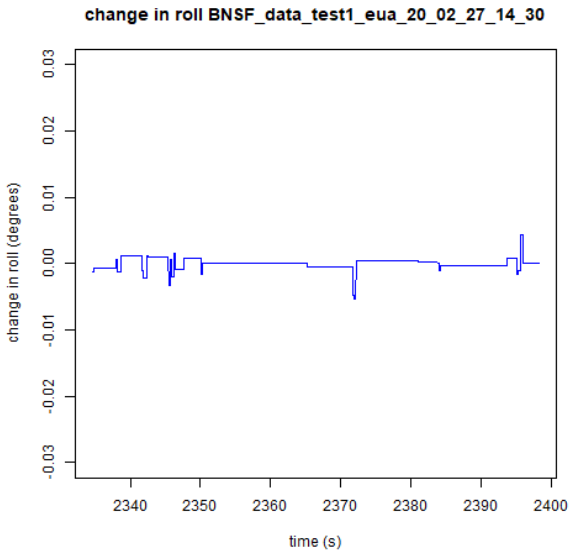


Figure 41. Differenced Roll.

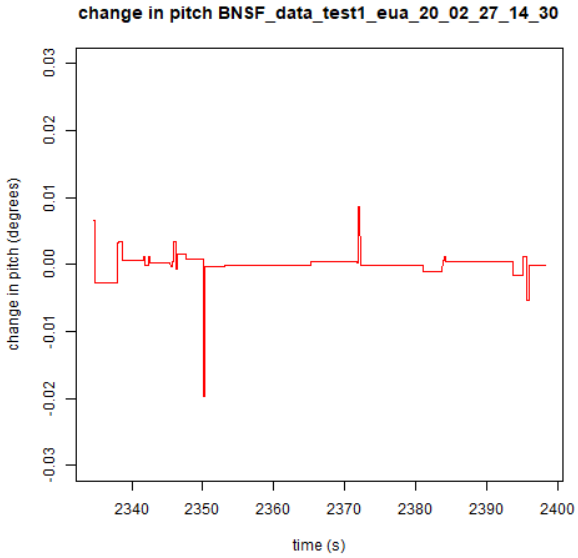


Figure 40. Differenced Pitch.

Peaks in the differenced Euler angle time series, which physically represent rotational velocity around the global track coordinates, were also investigated. After examining the differenced rotation time series, however, it was determined that creating an algorithm to detect their peaks would not be necessary nor would it be easy. Instead, Figure 41 and Figure 40 demonstrate that the peaks reveal themselves relatively clearly. Furthermore, significant changes in the SmartRocks orientation often occur around both the lateral and longitudinal global track axes simultaneously. While this EDA revealed interesting SmartRock behavior, indices were not calculated for the peaks found via an inspection of these plots, and thus the indices were not used in any additional analysis.

The last core calculated descriptive statistic to discuss in this section is cumulative rotation, the measure of the final orientation of the SmartRock versus the initial orientation after a train passing. This is simply found by taking the difference between the final and initial Euler angle in a Roll or Pitch time series, where Roll is the rotation around the lateral axis and Pitch is the rotation around the longitudinal axis, respectively per Figure 7, again not examining the Yaw due to those time series sometimes displaying unphysical behavior, varying in random patterns. Cumulative changes in a SmartRocks orientation after a train passes should correlate with the ballast dynamics experienced due to the train passing. This hypothesis was tested by comparing the calculated cumulative rotation values to various other calculated descriptive statistics to explore its validity. Figure 42 and Figure 43 show the cumulative rotation values plotted against the corresponding standard deviation of the same time series. While not tightly correlated, there is some observable relationship between these two statistics. The magnitude of the ratio between the two may be more specifically related to another categorical piece of information. Cumulative rotation, as well as its validity and relationship to other rotation and acceleration descriptive statistics, is explored in the Results chapter.

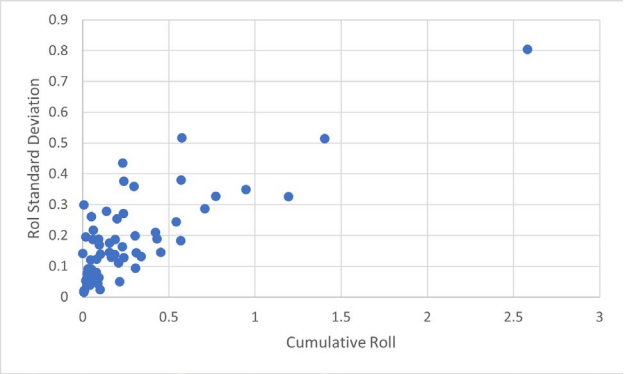


Figure 42. Standard dev. vs. cumulative Roll.

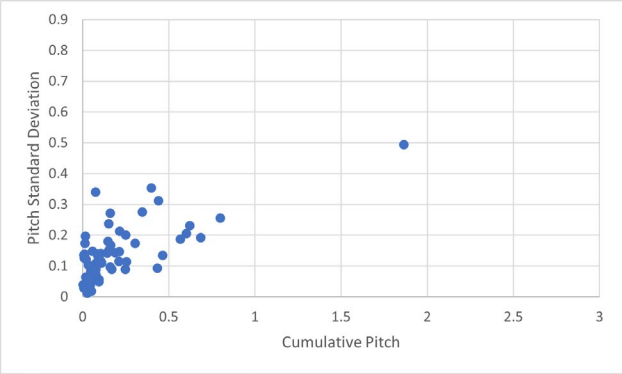


Figure 43. Standard dev. vs. cumulative Pitch.

DATABASE DETAILS

The data gathered for the purpose of this research report were recorded over a 2½-week period and were measured by a collection of 16 SmartRocks and 2 pairs of four strain gauge measurement devices. From this dataset, meaningful quantitative information about train passings recorded by the SmartRocks was extracted. Although data were first recorded on 2/1/2020, useful data were not recorded until 2/24/2020, after which only some of the day's SmartRocks recorded data properly recorded train passings, as Table 3 shows. After 3/9/2020, most of the SmartRocks' batteries began to run dry and data collection was no longer possible. For the datasets that did properly record train passings, the SmartRocks were only able to properly measure SmartRock acceleration and orientation. Three-dimensional stress was not properly measured due to improper calibration of the SmartRocks.

From the provided data, a consistent database was created that was organized by the categorical classifications discussed in this chapter. Afterward, MATLAB software was developed to visualize the data and perform a basic transformation on the acceleration data to put it in the global track coordinate system. Following this, from the acceleration time series, train passings were extracted and descriptive summary statistics were calculated to further visualize the calculated Euler angle time series and the acceleration time series.

From visualizations using distributions such as Q-Q and density plots, it was determined that the acceleration time series followed a Laplace distribution. Comparatively, Euler angle time series did not consistently follow a single distribution pattern. In particular, the rotation around the axis vertical to the track plane (the Yaw) did not follow any particular pattern. Often, the Yaw took the shape of Figure 20, but this was not always the case. For the most part, Roll and Pitch time series took similar shapes to the visualizations shown in Figure 22 and Figure 21, respectively, displaying more variation throughout any given train passing. As shown in this chapter, this behavior produced interesting descriptive statistics about the SmartRocks' movement throughout a train passing.

Additionally, an extensive EDA was performed to compare the calculated summary statistics across all 61 available train passings. From this, it was clear that quantitative differences exist across datasets and that they require further exploring to see what is fundamentally different between them. Using the descriptive summary statistics calculated in this chapter, a table of useful statistics was created to further compare the individual train passings using unsupervised machine learning methods fully discussed and explored in the following chapter. These summary statistics include:

1. SmartRock number, site, recording date, and location
2. Time series step size and data point count

3. Acceleration five-number summary
4. Acceleration peak counts
5. Euler angle five-number summary
6. Cumulative Roll and Pitch
7. Differenced Euler angle five-number summary
8. Train speed

The list above lays out all summary statistics calculated or gathered from the SmartRock data as discussed in this chapter. The summary statistics presented were used to calculate any additional quantitative measures (such as interquartile deviation). Most analyses introduced in the following chapter regarding differences between SmartRock behavior utilize the descriptive summary statistics calculated in this chapter. In addition to these explorations, additional EDA is explained that compares certain train passings more specifically.

Some other additional measures are introduced in the following chapter that were calculated from other descriptive statistics or measurements. An example of one of these is the time span associated with a train passing, calculated by multiplying the time step of the given time series by the number of data points. While not used extensively, this measure, and others like it, were used to compare differences in behavior between various datasets and how they may be due to different types of trains.

CHAPTER 4

Results

This chapter fully explores all analysis conducted into the SmartRock dataset beyond the exploratory data analysis discussed in the previous chapter. This includes unsupervised machine learning implementations used to diagnose differences in each SmartRock's behavior as well as a more detailed EDA geared toward identifying correlations between certain data parameters. From these analyses, numerous conclusions were drawn about the cause behind the differences seen in the SmartRock measurements. Lastly, a discussion section is included at the end of this chapter to summarize the results and explain their significance. Additionally, in this section, implications from these analyses as to how SmartRocks can be used to track ballast condition are discussed.

Most of the results presented in this research report were driven by the descriptive statistics introduced in the previous chapter. Due to the nature of the time series data recorded by the SmartRocks, many traditional time series analysis techniques are not conducive to achieving the goal of this research project. For example, time series forecasting techniques, which aim to predict what data values would follow after the time series concludes, are not useful in this scenario, since artificially extending acceleration or Euler angle signals, which is analogous to assuming a longer train, does not provide any information regarding ballast condition.

Furthermore, the size of the dataset made some analysis techniques not usable. As mentioned earlier, there were 61 SmartRock recordings, each of which contains three-dimensional acceleration and Euler angle time series. Hidden in each of these datasets is potentially useful information, which is extracted and explored in this chapter, but it is important to note that many of these recordings are of the same train by different SmartRocks. When accounting for this, the SmartRocks recorded only 11 train passings, 7 of which occurred at both testing sites. The relatively small number of recorded individual train passings proved to make certain machine learning techniques challenging to implement, since the ratio of data entries or variables in general should be kept large (Jenkins & Quintana-Ascencio, 2020). Unexpected results were encountered due to this challenge, which are discussed in this chapter.

Most analysis techniques implemented were quite successful, even with the presented challenges from the dataset. The secondary EDA includes all plotted data yielding interesting observations that have not been yet introduced. This includes plots of calculated summary statistics of all datasets across various combinations of time series to investigate their correlation. Correlation results between calculated descriptive statistics were also investigated using heat maps and correlograms. These helped identify which descriptive summary statistics should be compared.

K-means clustering algorithms were implemented to cluster datasets in all tertiary EDA results that had unique characteristics. Clustering was the primary method used to diagnose differences between datasets based on what cluster they belonged to in various summary statistic comparison plots. From these results, the effect of each qualitative dataset classifications was identified, illustrating clearly what is the likely cause of differences in datasets.

The attempted implementation of various regression modeling techniques to verify observed differences between datasets is fully discussed. The primary result is a logistic regression analysis to identify differences in SmartRock behavior based on their installation site. This section also includes regression

modeling attempts such as multi-adaptive regression spline, which did not yield as interesting of results due to challenges experienced with the dataset.

Finally, other analysis techniques that were hypothesized and attempted throughout this research project are discussed. The challenges that arose while attempting to implement these techniques are also presented. This section includes all analysis techniques that did not necessarily provide new quantitative information of the SmartRock behavior but did help guide further research explorations, such as Fourier analysis of SmartRock response data.

This chapter concludes with a discussion of all presented results and their implications on substructure dynamics. This includes a broad discussion of the analysis results as a whole and what they imply about the capabilities of SmartRock as a ballast dynamics measurement hardware. Any unexpected results yielded throughout this research project are also discussed.

SECONDARY EDA

This section explores all additional informative EDA performed. These analyses are included here, since they are more descriptive and directly inform the remaining analyses in the chapter. The EDA in this chapter looks at the whole SmartRock dataset but also investigates various subsets of the dataset based on the qualitative classifications presented in the previous chapter.

A correlogram, shown in Figure 44, was calculated to investigate how various variables, in the form of descriptive summary statistics, correlate with one another. The size of the bubble in the matrix indicates the level of correlation between two variables, and the color indicates positive (blue) and negative (red) correlation based on the Pearson correlation coefficient (Pearson, 1895). The equation on the following page shows how the correlation coefficient r is calculated for two sets of samples x and y of length n. The correlation coefficients displayed in the correlogram were calculated using the *cor()* function in R (Wei & Simko, 2017).

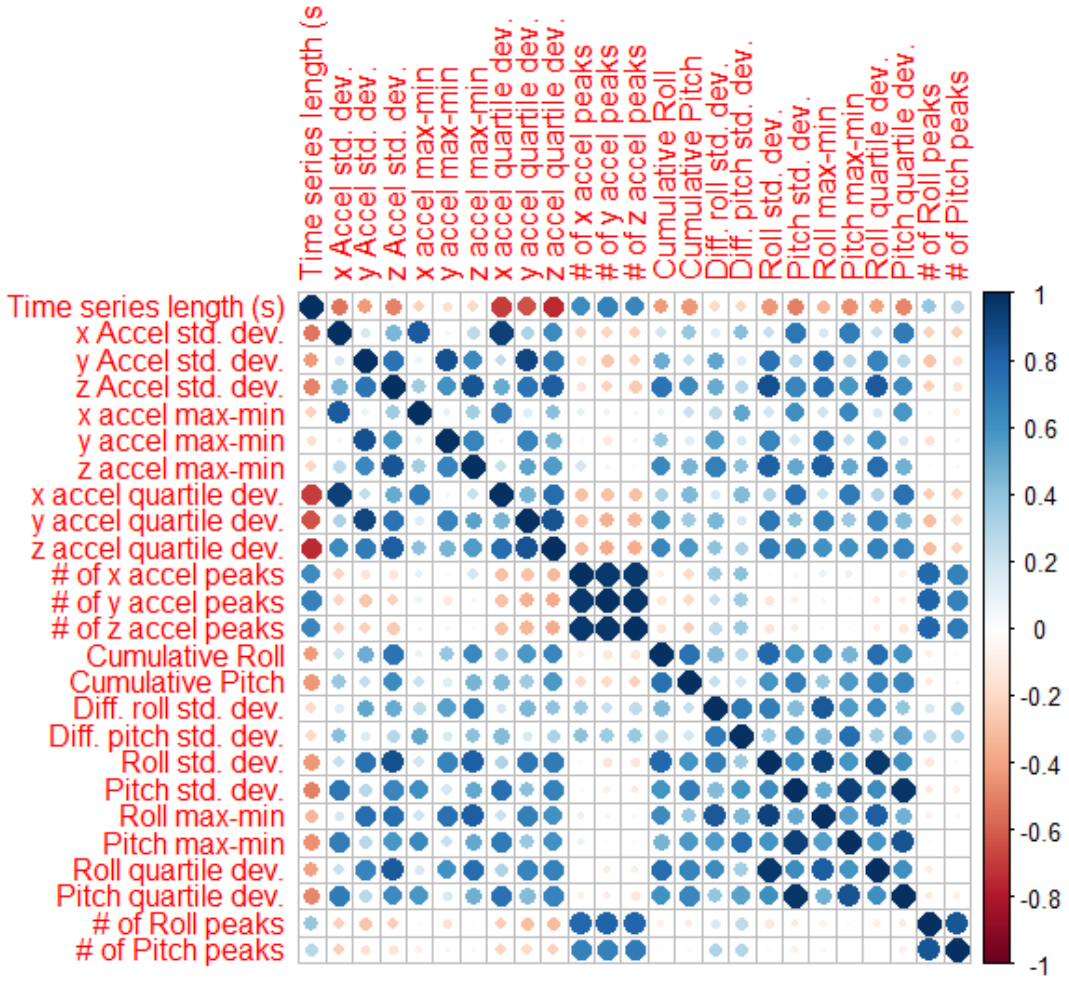


Figure 44. Correlogram comparing all calculated SmartRock statistics.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Where

r_{xy} = correlation coefficient
x,y = sets of samples of length n with i indices

Most of the variables in the correlogram were previously explained. The time series length is the time step of a given time series multiplied by the number of data points in the time series, yielding an estimate of the time it took for the train to pass over the track site. Expectedly, the time series length correlates well with the count of acceleration peaks in all three dimensions, indicating that longer trains impose more load impulses, which likely represent peaks in the acceleration time series. Furthermore, the time series length is inversely correlated with the acceleration quartile deviations, indicating that longer train passings result in smaller interquartile ranges. The acceleration quartile deviations plotted against the time series length are shown in Figure 45, with squared residuals shown to compare to the correlation values displayed in the correlogram.

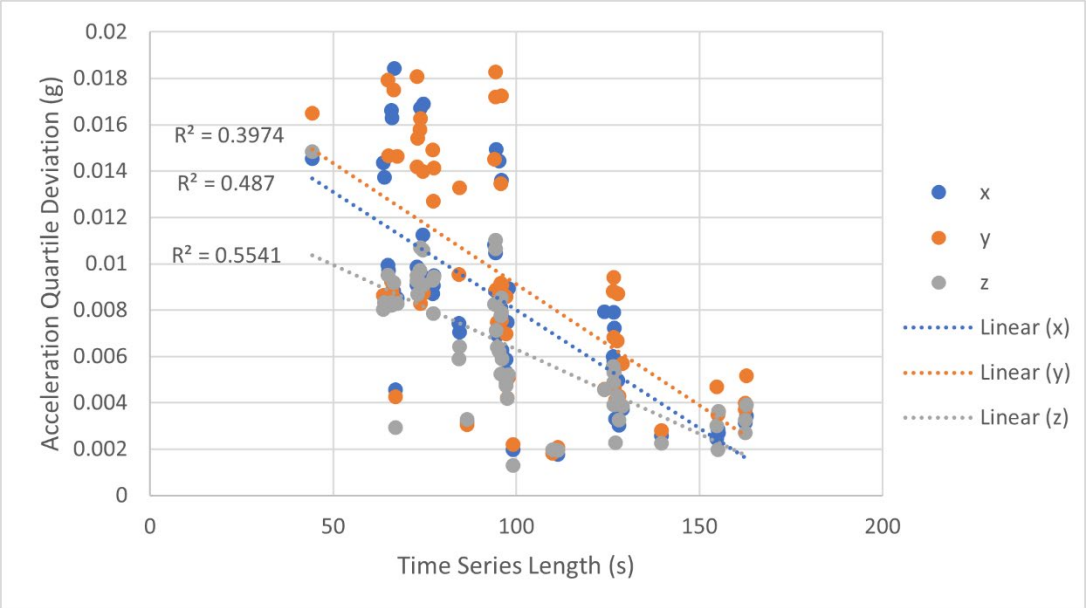


Figure 45. Acceleration quartile deviations vs. time series length.

Beyond correlations with the time series length, other correlations illustrated in Figure 44 are important to note. Some were expected, such as the correlation between acceleration max−min and acceleration quartile deviations, as already shown in Figure 35. From this, it was also expected that acceleration standard deviations would correlate with acceleration ranges and quartile deviations. Comparisons of acceleration standard deviations against quartile deviations are shown in Figure 46. As the correlogram suggests, these variables are highly correlated. It is also important to note that the magnitude of these variables was larger in the x and y dimensions than the z dimension. This observation is further explored later.

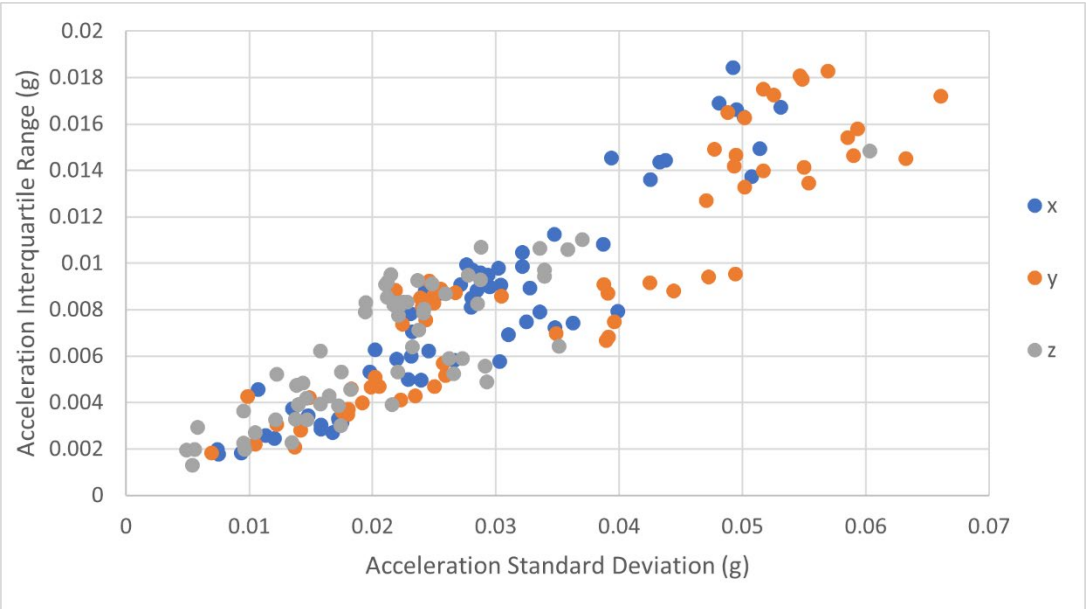


Figure 46. Acceleration quartile deviations vs. standard deviations.

Another interesting pair of correlating variables were acceleration and Euler angle descriptive statistics across dimensions. For example, y acceleration standard deviation correlates well with Roll standard deviation while x acceleration standard deviation correlates well with Pitch standard deviation. This makes sense, since Pitch is the rotation around the axis longitudinal to the rail which lateral forces drive. Correlation of summary statistics calculated from these pairs of time series are seen throughout the correlogram. Figure 48 and Figure 47 show the linear correlation of rotation and acceleration standard deviations across dimensions. From these plots it is also clear that variations in longitudinal forces imposed on the SmartRocks, and the resulting variations in rotation around the lateral track axis (along the tie), are larger than variations due to lateral forces.

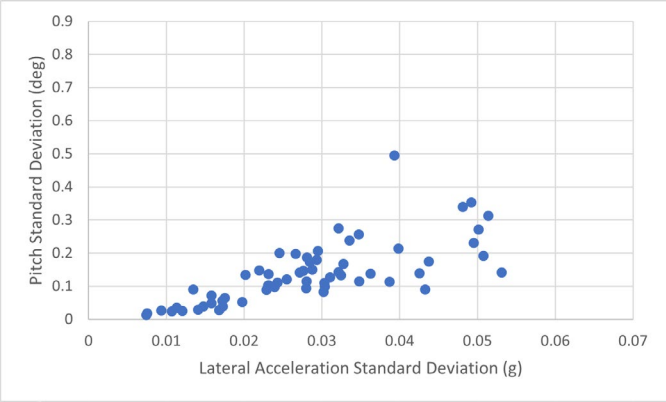


Figure 48. Pitch vs. lateral acceleration.

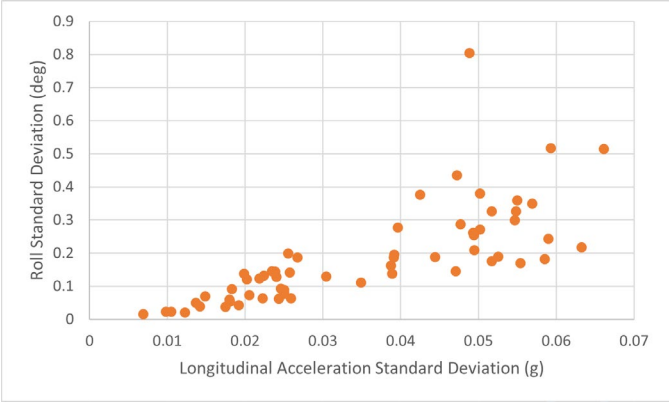


Figure 47. Roll vs. longitudinal acceleration.

Continued observations of the correlogram also revealed the z acceleration variables correlate with both Roll and Pitch rotation variables. This can be interpreted as larger variations in vertical forces imposed on the SmartRock, generally resulting in larger variations in SmartRock rotation. Figure 49 shows Roll and Pitch standard deviation values plotted against vertical acceleration standard deviation values. Both pairs of variables correlate relatively well, with Roll and vertical acceleration standard deviations showing better correlation. Furthermore, this pair of variables linear best fit has a larger slope than that of Pitch and vertical acceleration standard deviations.

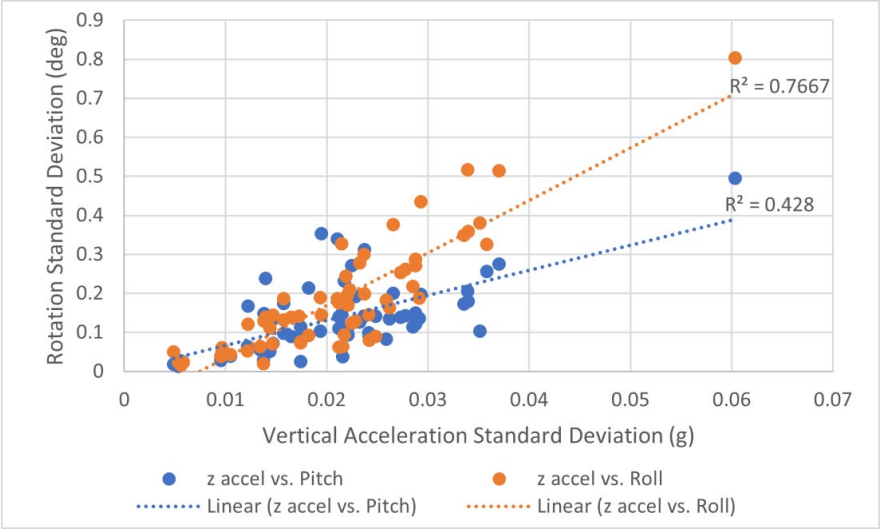


Figure 49. Roll and Pitch vs. vertical acceleration standard deviations.

The correlation of various variables was additionally explored by plotting a Heat Map of the variables using the *heatmap()* function in R (R Core Team, 2020; Wei & Simko, 2017). The dendrograms in Figure 50 are calculated by applying agglomerative hierarchical clustering to the same set of descriptive statistics previously explored. Hierarchical clustering orders clusters based on the similarity of data vectors, which is illustrated by dendrograms (Everitt, 1998). Pearson correlation values were fed into a complete-linkage (or farthest neighbor) hierarchical clustering algorithm to produce the dendrograms (Sorensen, 1948).

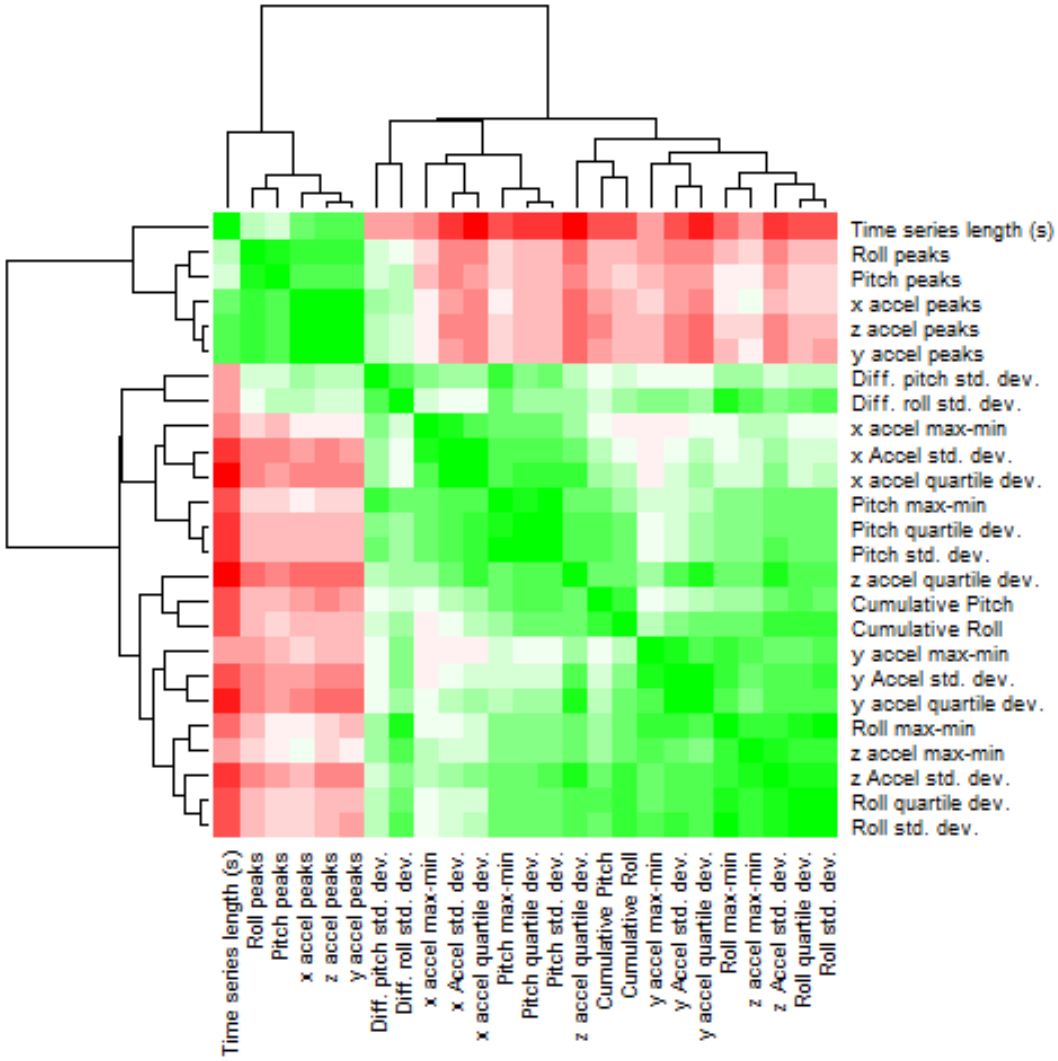


Figure 50. Heat Map comparing all calculated SmartRock statistics

At each step of the agglomerative complete-linkage clustering algorithm, the two clusters separated by the shortest distance are combined. Using this algorithm, the link between two clusters contains all element pairs and the distance between clusters equals the distance between those two elements that are farthest away from each other. Mathematically, this process is described by the equation below. *D(X,Y)* is calculated for every unique cluster combination, with minimum values being hierarchically clustered first, as illustrated by the dendrogram.

$$D(X,Y) = max_{x \in X, y \in Y}\, d(x,y)$$

Where

$D(X,Y)$ = distance between clusters X and Y
d(x,y) = Euclidean distance between element-wise pairs where $x \in X, y \in Y$

As the dendrograms in Figure 50 show, the SmartRock descriptive variables were broadly clustered into two separate clusters, one which contains information on the time series length and number of peaks and one which contains calculated summary statistics. Working down the right arm of the top cluster, two additional cluster arise, the left one predominantly composed of statistics calculated from x acceleration and Pitch time series. These results support findings proposed earlier in this section. The dendrogram also reinforces that z acceleration standard deviation values are highly correlated with Roll quartile deviation and standard deviation values.

A few new correlations are also revealed in the dendrogram. First, absolute cumulative Roll and Pitch values are most closely associated with z acceleration quartile deviation values rather than x or y acceleration summary statistics. Figure 51 shows these variables plotted against one another. The initial reaction to this result is that large variations in a SmartRock's vertical acceleration, resulting likely from dynamic wheel impacts large in magnitude, are associated with above-average cumulative changes in a SmartRock's orientation. This correlation is explored later in this chapter.

This section showed that numerous SmartRock descriptive statistics are correlated with one another in both expected and unexpected ways. The observed results also show key differences between dimensions of acceleration and Euler angle time series. Since these plots investigate all SmartRock recordings collectively, there are 87 numerous underlying factors that could potentially be attributed to causing these differences. The remainder of this chapter investigates these underlying causes by exploring unsupervised machine learning applications on the time series and calculated descriptive statistics.

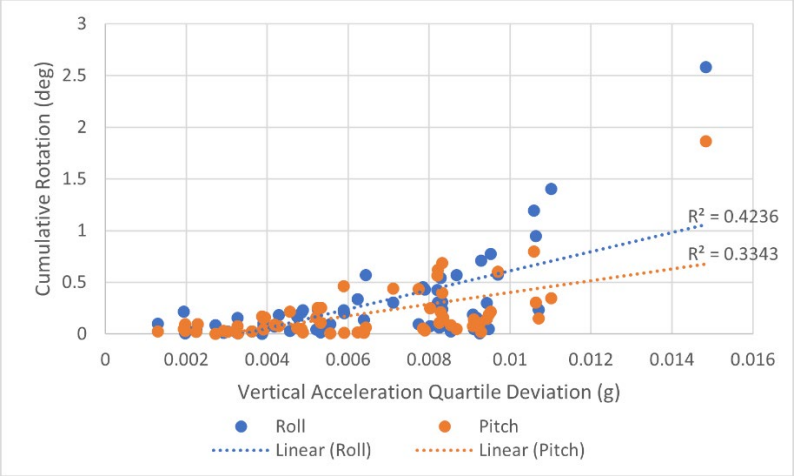


Figure 51. Abs, cumulative Roll and Pitch vs. vertical accel. quartile deviations

K-MEANS CLUSTERING

K-means clustering is a signal processing method of vector quantization that partitions n observations into k number of clusters, determined by the user, with the closest centroid, minimizing within-cluster variances (MacQueen, 1967). Manually calculating these centroids is computationally difficult, even for only 61 samples, but efficient heuristic algorithms converge to local optimal centroids relatively quickly. The *kmeans()* function in R was used to produce the results that follow in this section (R Core Team, 2020). The Hartigan-Wong variation of the k-means algorithm, which progresses toward a local minimum of the

minimum sum-of-squares of samples from their cluster's centroid, was utilized (Hartigan & Wong, 1979). This method is classified as a local search that iteratively attempts to change the cluster classification of a sample, keeping it in its new cluster if the sum-of-squares objective is reduced. Initially, each sample is assigned to a random cluster $\{S_j\}_{j\in\{1,\cdots,k\}}$. Updates are calculated iteratively using the equation below. For the set of $\{x, n, m\}$ that the equation below reaches a maximum, x moves from cluster S_n to the cluster S_m. The algorithm is completed once $\Delta(m,n,x)$ is less than zero for all sets of $\{x, n, m\}$. Once complete, the set of samples are clustered into user-defined k clusters $\{S_j\}_{j\in\{1,\cdots,k\}}$ with the minimum sum-of-squares of sample's euclidean distance for their cluster's centroid (Hartigan & Wong, 1979).

$$\Delta(m,n,x) = \phi(S_n)\phi(S_m) - \phi(S_n\backslash\{x\}) - \phi(S_m \cup \{x\})$$

Where

$\phi(S_j)$ = individual cost of S_j defined by $\sum_{x\in S_j}(x = \mu_j)^2$

μ_j = centroid of cluster j

n,m = cluster indices

K-means clustering was used to partition SmartRock datasets based on two descriptive statistics plotted against one another. This method builds upon the EDA done in this chapter and reveals differences between the SmartRock datasets. Before applying this algorithm to the SmartRock data on a wide scale, the number of clusters necessary to minimize the total within-clusters sum-of-squares had to be determined. When the number of clusters is equivalent to the number of samples, the sum-of-squares is zero, since each sample has the same value as its cluster's centroid. Conversely, one cluster maximizes the sum-of-squares. A happy medium between these two extremes must be found to best characterize the datasets using k-means clustering.

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS

The following analysis shows how the optimal number of clusters was determined. This was done using numerous pairs of summary statistics to see if the optimal number of clusters varied based on which values were used. It was determined that this was generally not the case, as this section shows. Figure 52 shows the total within-cluster sum-of-squares for various numbers of clusters in the form of an elbow diagram. As expected, the sum-of-squares reduces with increasing number of clusters k. The ideal number of clusters is the number after which the total-within sum-of-squares does not decrease significantly (Ketchen & Shook, 1996). In this case, the optimal number of clusters is 5, after which the sum-of-squares does not decrease much further. Figure 53 show the k-means clustering results for values of $1 \leq k \leq 6$, where the solid black triangles are the cluster centroids. Based on inspection, Figure 53(f) looks very reasonable, but further improvements could be made.

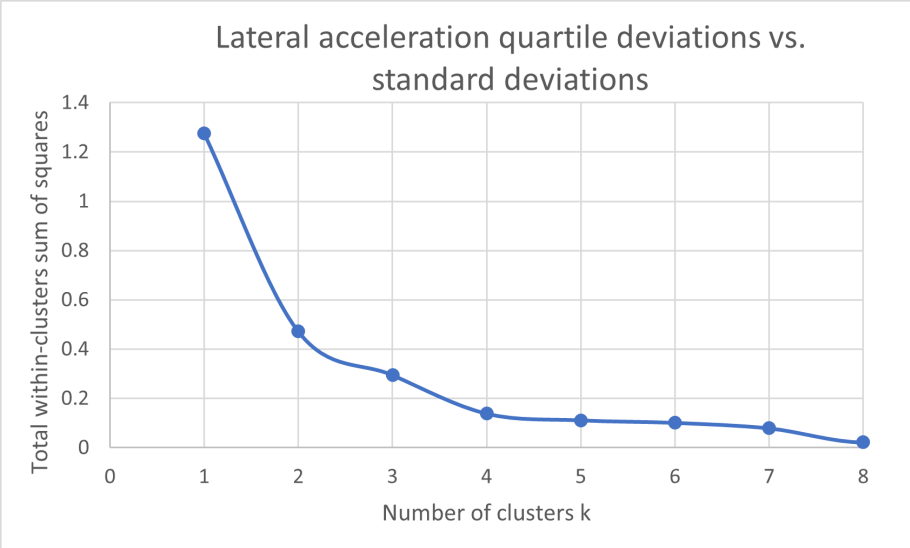
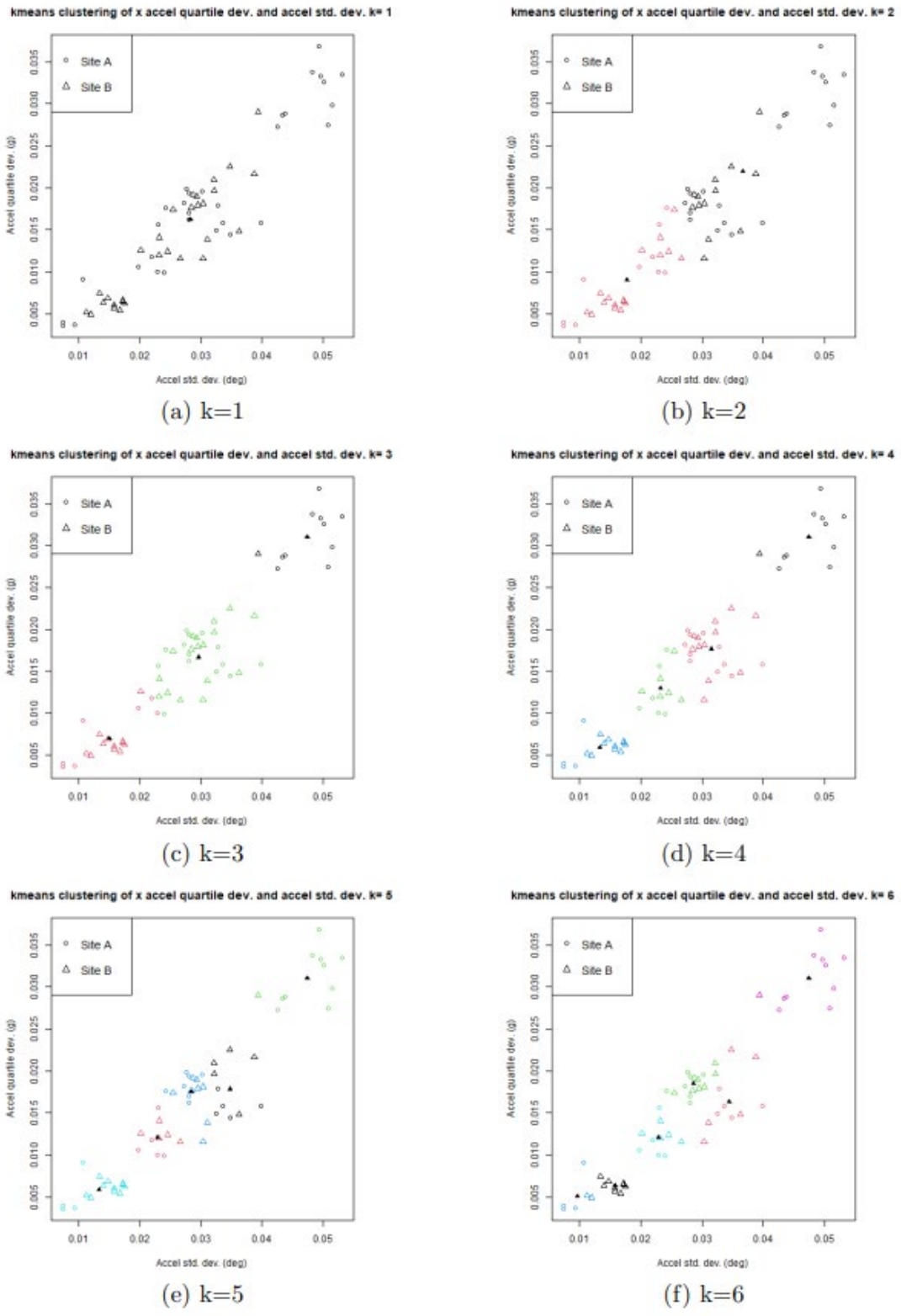


Figure 52. Total within-cluster sum-of-squares.



(a) k=1 (b) k=2

(c) k=3 (d) k=4

(e) k=5 (f) k=6

Figure 53. K-means clustering results based on value of k.

While quantitatively determining the number of clusters is a good method of verifying that the total-within sum-of-squares is minimal, most k-means results in this report were analyzed based on inspection rather than elbow diagrams. Most elbow diagrams calculated for k-means results quantitatively determined that the optimal number of clusters was typically k = 5 or k = 6. The remaining k-means results introduced in this chapter illustrate that this number of clusters could be reasonably assumed.

K-MEANS RESULTS

K-means clustering results were the first analysis technique that allowed quantitative differences in SmartRock datasets to be linked to qualitative differences. Most analysis initially investigated the vertical acceleration descriptive statistics and how they correlated with various rotation and acceleration summary statistics. Figure 55 and Figure 54 show k-means clustering results for vertical acceleration standard deviations against lateral and longitudinal acceleration standard deviations, respectively. Initial observations of these results show that there is a general trend between these two statistics, indicating that large variations in SmartRock vibration in one dimension may indicate SmartRock vibration in the other dimensions. Furthermore, acceleration standard deviation values tend to cluster together based on their installation site, as illustrated by the chosen symbology. Lastly, these images show that lateral acceleration standard deviation values tended to be higher than longitudinal ones.

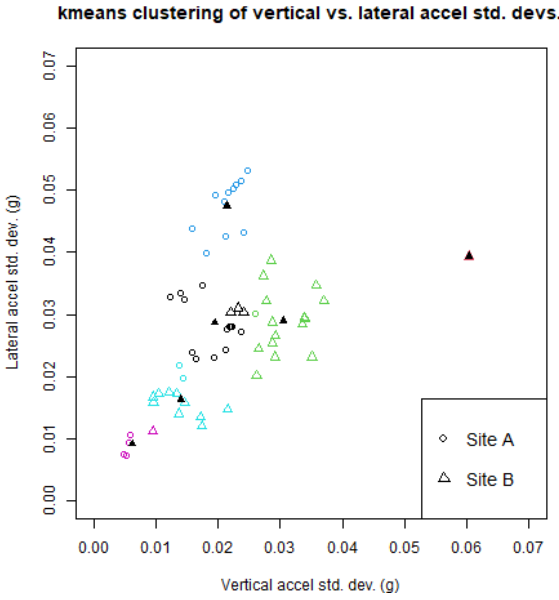


Figure 55. x vs. z.

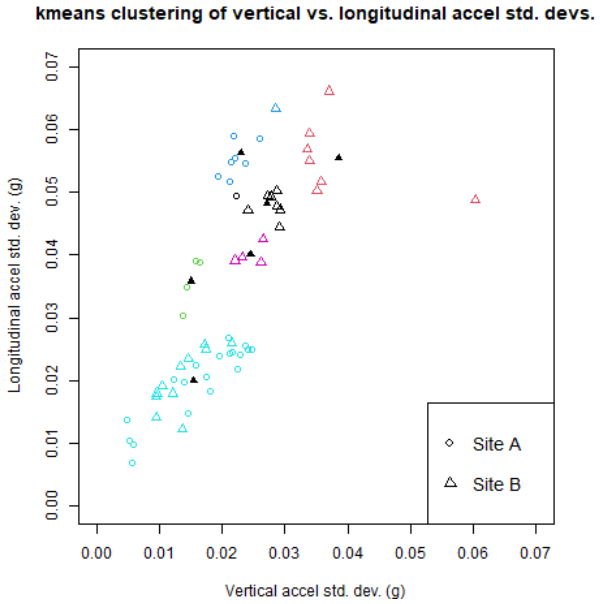


Figure 54. y vs. z.

This section is divided into subsections to describe specifically the different k-means clustering results calculated and their implications. The first subsection discusses observed correlation between lateral and longitudinal acceleration summary statistics and potential reasons for them. The second subsection investigates correlation of the calculated cumulative rotation summary statistic calculated for both the Roll and Pitch series and their implications.

LATERAL VS. LONGITUDINAL ACCELERATION

Interesting observations were made when comparing lateral and longitudinal acceleration summary statistics. Figure 57 and Figure 56 show the lateral and longitudinal quartile deviation values plotted against

one another for two different groups of SmartRocks. Observations of the corresponding datasets of each value in the plot led to the realization that there appeared to be two separate patterns forming, depending on which SmartRocks were considered, and they were plotted as such. For SR1, SR2, and SR6 the x quartile deviation is larger than the y quartile deviation, while the opposite is observed for the rest of the SmartRocks. It is important to note that these two groups were not split by installation site, however, as SR7 and SR8 are part of Site A, which includes under-tie pads.

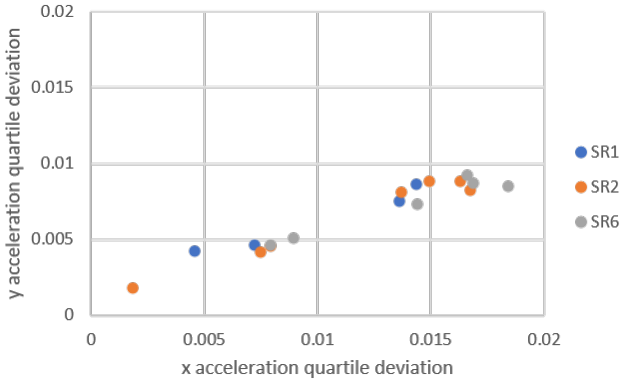


Figure 57. SmartRock Group 1.

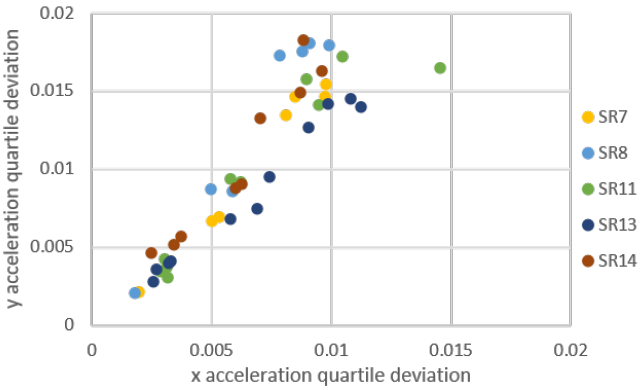


Figure 56. SmartRock Group 2.

It was also found that the recording date of these summary statistics correlated as well, as shown in Figure 59 and Figure 58, which compare both lateral and longitudinal acceleration quartile deviation and standard deviation values, respectively. Both plots show correlation of values based on the recorded train passing. Furthermore, especially in Figure 59, distinct groups appear to be present. This observation is an example of why k-means clustering was determined to be a useful analysis technique.

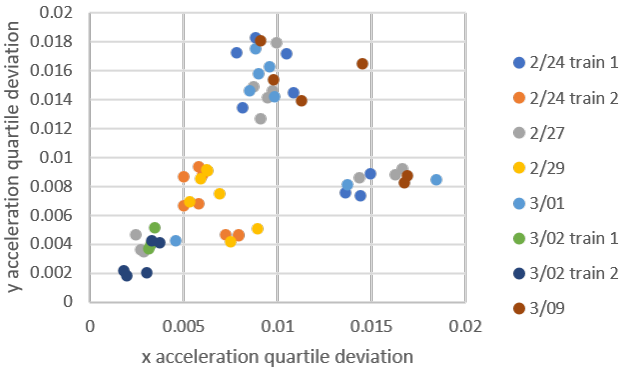


Figure 59. Quartile Deviation.

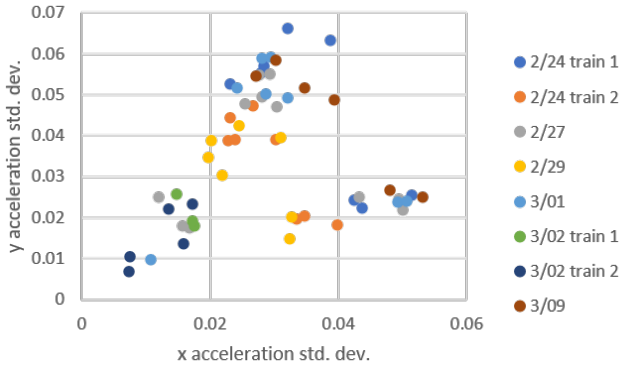


Figure 58. Standard Deviation.

The number of clusters used for the following analysis was k = 5, although using six would have also been valid, since an outlier appears to be present in Figure 59. Figure 60 shows k-means clustering performed on the lateral and longitudinal acceleration quartile deviation values shown in Figure 59, with a legend included to detail the symbols chosen. Table 8 shows the nomenclature used for discussion of these clusters throughout this section.

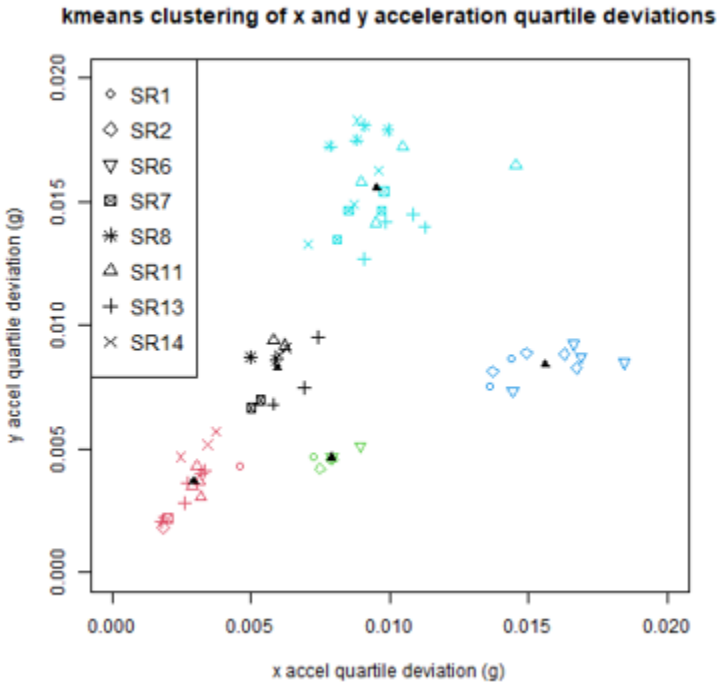


Figure 60. Clustering of lateral and longitudinal accel. quart. deviations by SR num.

Table 8. Clusters and corresponding colors for results.

Clusters:

- #1 - Black
- #2 - Red
- #3 - Green
- #4 – Navy Blue
- #5 - Teal

Various k-means clustering of these values was performed with different symbologies included to demonstrate how datasets were correlating with other datasets of similar or different characteristics. The k-means clustering algorithm for this dataset is well-behaved, and clusters form as expected based on inspection. Since determining which cluster various SmartRock datasets belong to is difficult strictly based on inspection, a bar chart was created to quantify which clusters SmartRock datasets were falling into, shown in Figure 61. This technique was explored after earlier k-means clustering analysis results, which obscured information about the dataset. This plot reaffirms earlier observations that two groups of SmartRocks were falling into two distinct trends. If Cluster 2 (red) is not considered, which contains datasets with small calculated lateral and longitudinal acceleration quartile deviation values, every SmartRock either belongs to both Clusters 3&4 (green and blue) or 1&5 (black and teal). The colors used to identify each cluster are the same throughout this analysis. Since these two groups of SmartRocks differences cannot be attributed to site of installation, it is difficult to determine why distinct trends are arising from the data.

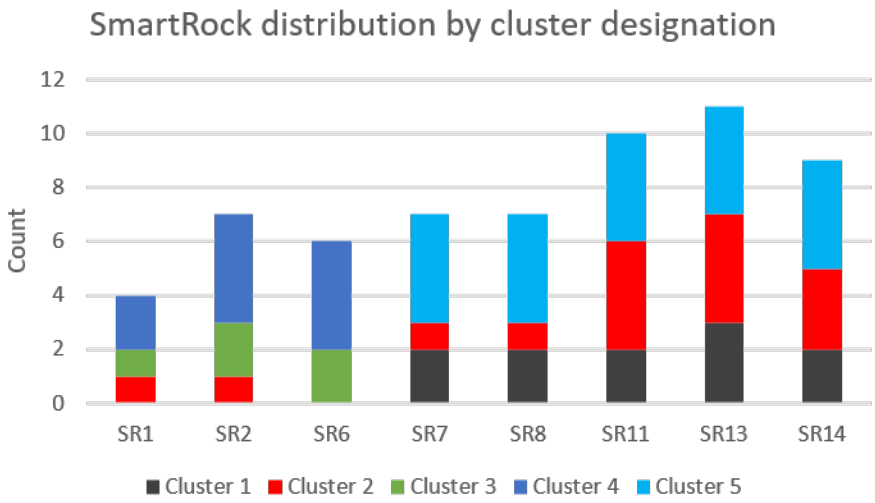


Figure 61. Bar chart showing what clusters SmartRocks belong to by number.

These noted trends of SmartRock datasets using lateral and longitudinal acceleration quartile deviations are supported by Figure 62, which shows lines of best fit for each SmartRock's collection of datasets, with five SmartRocks having a distinctly steeper trend than the other three. Furthermore, these lines of best fit are highly correlated. Table 9 shows the average absolute residual for each of the lines of best fit, which are sufficiently small for the trendline to be considered appropriate.

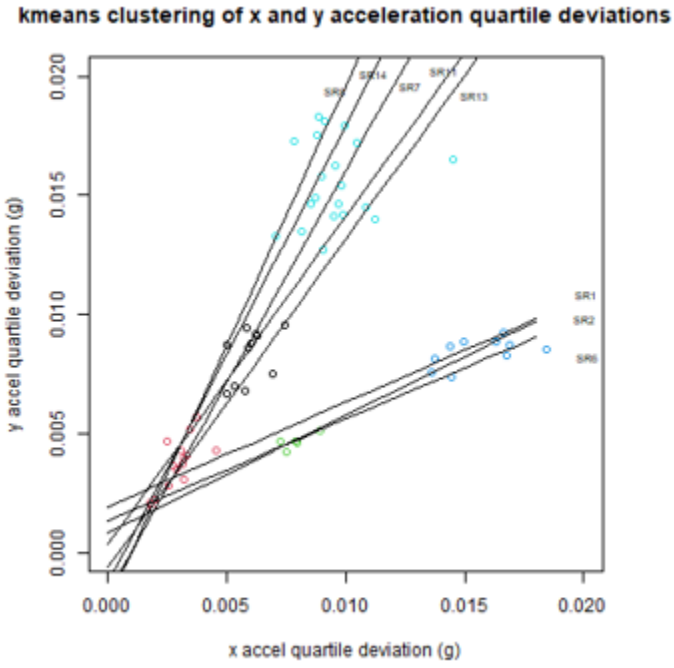


Figure 62. Longitudinal vs. lateral accel. quart. deviations by SR number.

Table 9. Average absolute residuals for each line of best fit shown above.

SR #	Average Absolute Residual
1	3.78E-04
2	3.72E-04
6	3.25E-04
7	6.82E-04
8	1.03E-03
11	1.51E-03
13	5.63E-04
14	1.08E-03

Analogous k-means clustering analysis was performed based on the location of the SmartRocks with respect to the rails and ties to support the findings shown from analysis regarding SmartRock numbers. As discussed earlier in this report, SmartRocks were installed in four distinct locations: in front of vs. behind the tie and gauge side (GS) vs. field side (FS) of the track. Figure 63 shows the same k-means clustering results, with SmartRock location denoted in the legend, and Figure 64 illustrates how the SmartRock location affected cluster identity. The takeaway from this analysis is that Cluster 4 only included SmartRocks that were installed in front of the tie, which were SR2 and SR6. Quantitatively speaking, SmartRocks installed in front of the tie experienced larger variations in lateral forces than longitudinal ones.

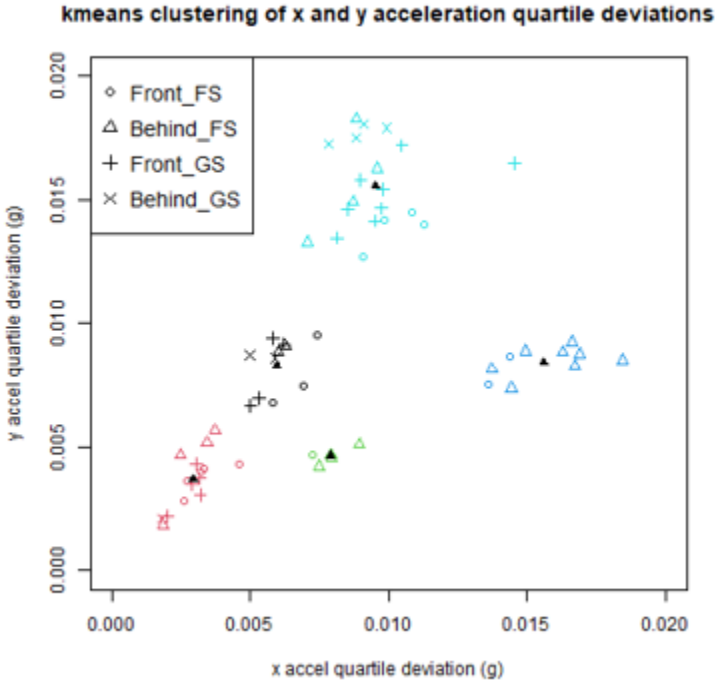


Figure 63. Clustering of lateral and longitudinal accel. quart. deviations by SR loc.

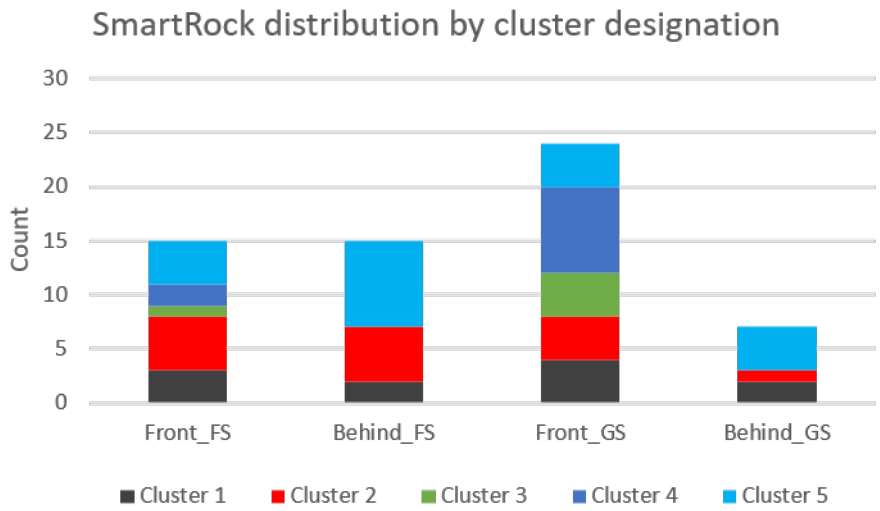


Figure 64. Bar chart showing what clusters SmartRocks belong to by location.

Finally, the same analysis was performed focusing on the recorded train passings by date for all the SmartRock datasets. This was done to determine if different trains passings influenced the calculated acceleration quartile deviation statistics. There were a few notable observations made from Figure 65, which displays the cluster to which each dataset belongs by recorded train passing per Figure 66. First, all SmartRock datasets recorded on 3/02/2020 were part of Cluster 2, meaning their acceleration quartile deviation values were small. This likely means that the train passing recorded on this date was of a lighter axle load. Next, datasets that were calculated to be a member of Cluster 1 were predominantly recorded on 2/24/2020 (train 2) or 2/29/2020. This could indicate some similarities between these two train passings. Lastly, all four datasets that were determined to be a member of Cluster 3 were determined to be recorded on either 2/24/2020 (train 2) or 3/09/2020. Since this cluster only contains SR2 and SR6, as determined earlier, this result further enforces similarities between them. This is somewhat unexpected, since both SmartRocks were installed at the same location with respect to the rail and tie at Site A, meaning their measurements should be relatively similar.

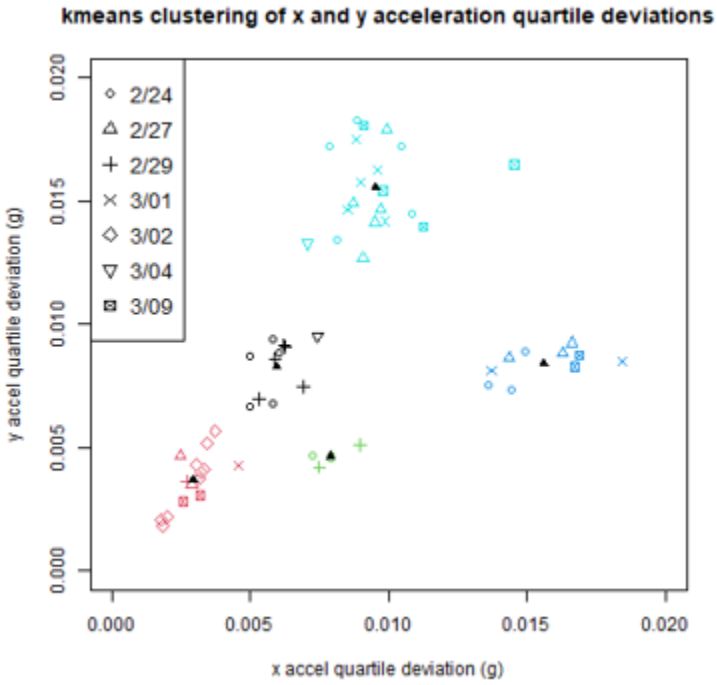


Figure 66. Clustering of lateral and longitudinal accel. quart. deviations by SR date.

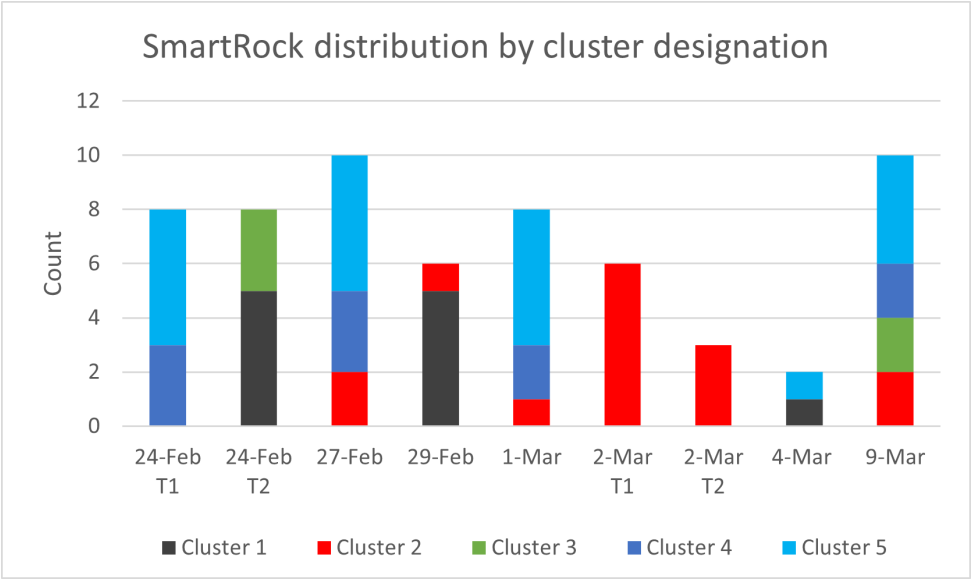


Figure 65. Bar chart showing what clusters SmartRocks belong to by recording date.

Knowing that the lateral and longitudinal acceleration standard deviations also correlated linearly with vertical acceleration standard deviations, a 3D plot was made to compare all three acceleration dimensions' quartile deviations, shown in Figure 67 with data points colored analogously to the clusters calculated and investigated throughout this analysis. From this plot, we see that two distinct trends are still present. Both lateral and longitudinal acceleration quartile deviations linearly correlated with vertical acceleration quartile deviations similarly.

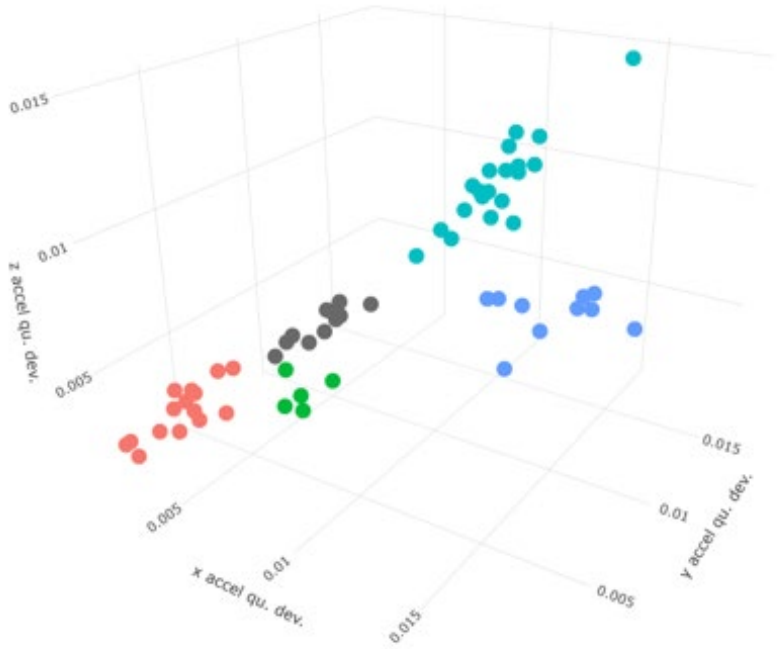


Figure 67. 3-D plot showing all three acceleration dimensions' quartile deviation values.

CUMULATIVE ROTATION

Cumulative rotation was introduced as a descriptive statistic in this analysis to investigate how SmartRocks angularly displaced throughout a train passing. The validity of this summary statistic was first verified by simply plotting absolute cumulative rotation against rotation standard deviations. It was expected that larger angular vibrations in a SmartRock would likely cause more cumulative rotation. This was verified in Figure 68 and Figure 69, which show a general linear trend between these two variables. These plots also show that SmartRocks were capable of notably displacing angularly, even though the corresponding standard deviation was small. Lastly, these plots provided the first evidence that cumulative Roll was larger than cumulative Pitch in general.

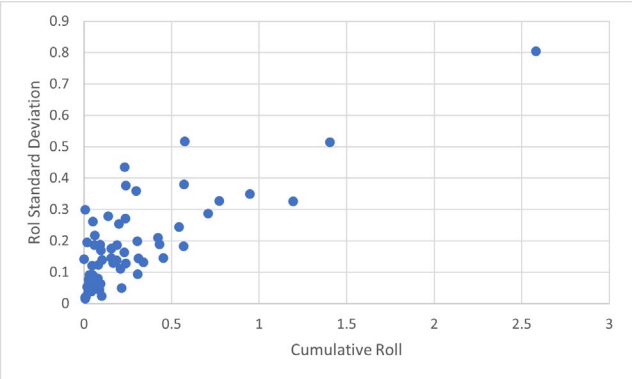


Figure 68. Roll statistics

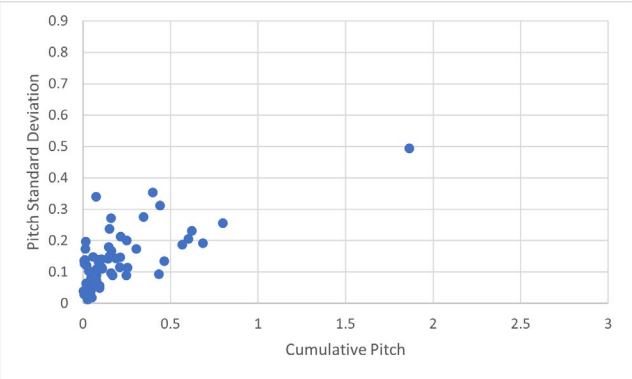


Figure 69. Pitch statistics.

A similar analysis was done for Yaw summary statistics. Figure 70 shows k-mean clusters calculated for Yaw summary statistics and shows the cumulative Yaw versus Yaw standard deviation for all datasets. From this plot it appears that these summary statistics are highly correlated. This observation led to further investigations into the Yaw time series, which showed odd behavior in many cases. Thus, although interesting, this particular result was not used to make any conclusions on the SmartRock's ability to quantify ballast dynamics.

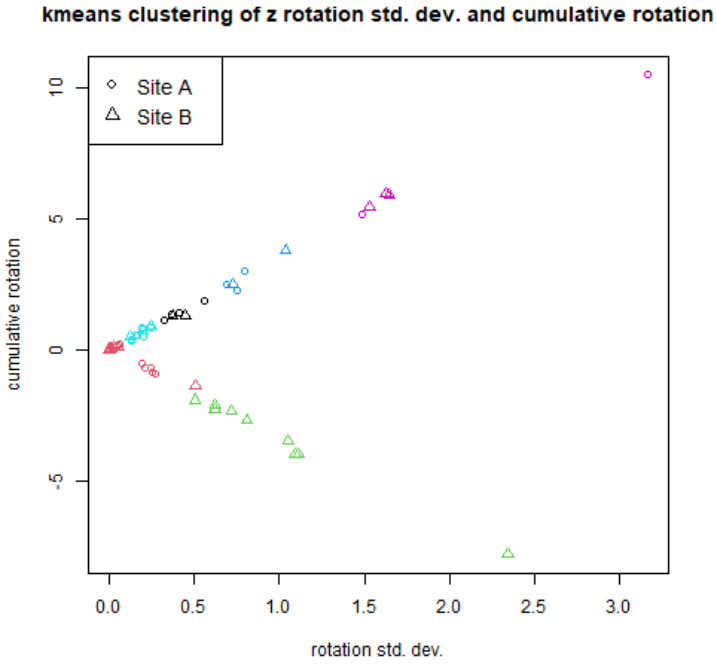


Figure 70. K-means clustering of Yaw statistics.

Vertical acceleration standard deviation values were evaluated against rotation summary statistics as well. The results, shown in Figure 72 and Figure 71, are relatively like what was seen in the previously introduced k-means clustering plots. Again, SmartRocks installed at the same site appear to be tightly knit but don't necessarily belong to the same calculated clusters from the Hartigan-Wong method. Even though the one outlier vertical acceleration standard deviation value is still present, it is part of the general linear trend seen in these plots. Furthermore, Roll standard deviation values tended to be larger than Pitch standard deviation values, which supports earlier findings. While these results are interesting, none of them were directly obtained from the k-means clustering algorithm. Since all datasets were included, it is difficult to determine how categorical differences in the datasets were driving cluster determination from these plots alone. Similar observations were drawn from Figure 74 and Figure 73.

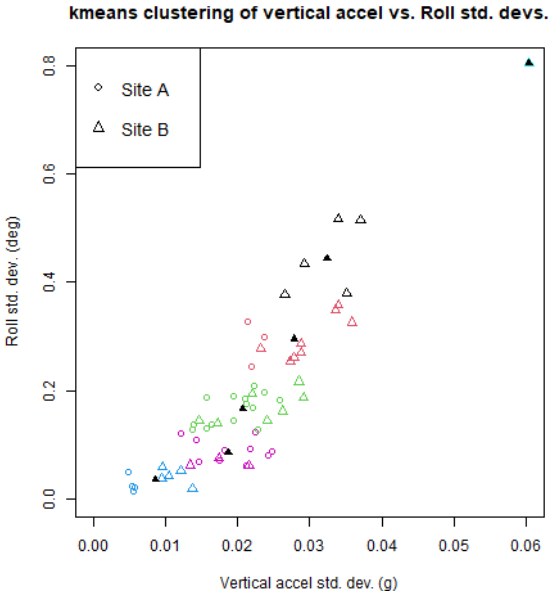


Figure 72. Roll vs. vert. accel. s.d.

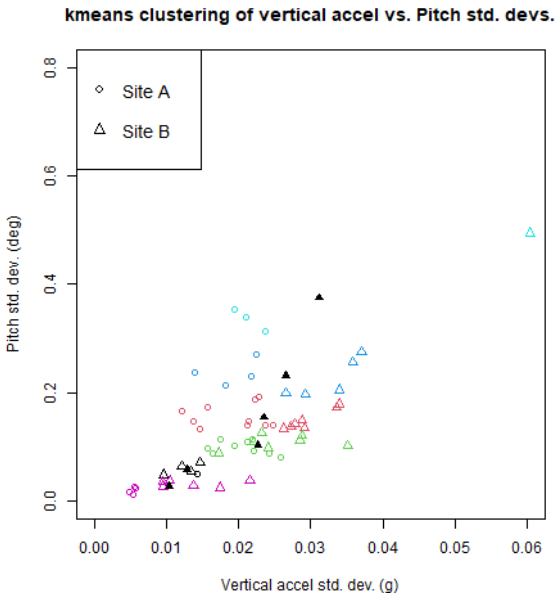


Figure 71. Pitch vs. vert. accel. s.d.

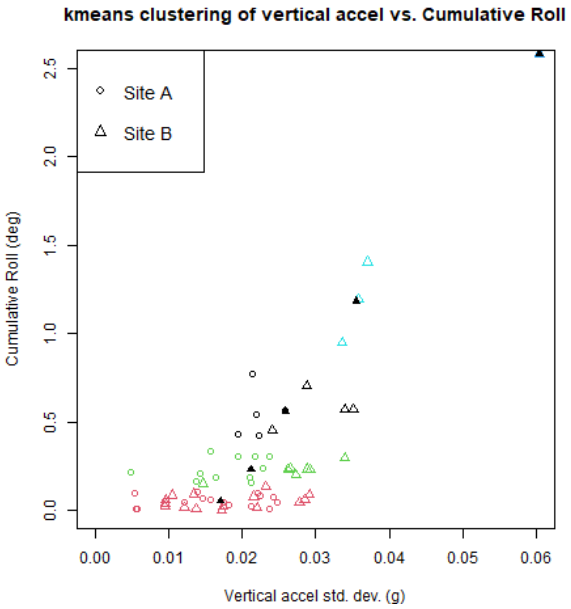


Figure 74. Cum. Roll vs. vert. accel. s.d.

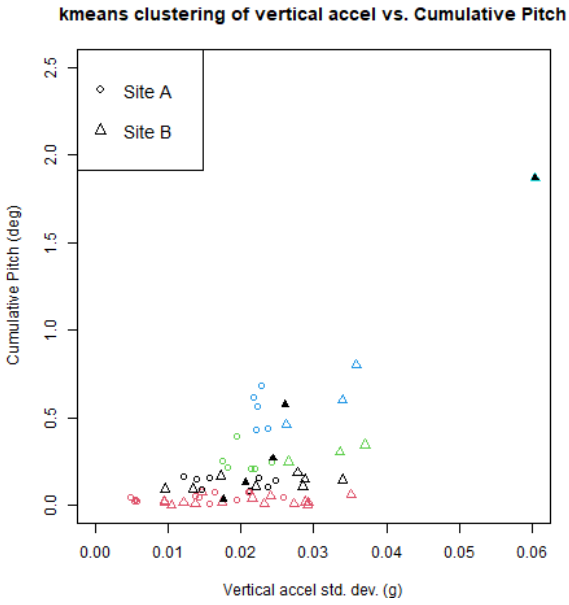


Figure 73. Cum. Pitch vs. vert. accel. s.d.

Correlations of cumulative Roll values, when plotted against vertical acceleration standard deviation, were better illustrated when plotting SmartRocks for each installation site individually, as shown in Figure 76 and Figure 75. These plots show that both cumulative Roll and vertical acceleration standard deviation statistics were generally larger in magnitude for Site B SmartRocks than for Site A SmartRocks. Several observed cumulative Roll values were quite large for Site B SmartRocks compared to a maximum of 0.8 between all Site A SmartRocks, indicating significant movement of Site B SmartRocks under some passing trains. This observation shows a clear effect of the under-tie pads at Site A on mitigating SmartRock movement.

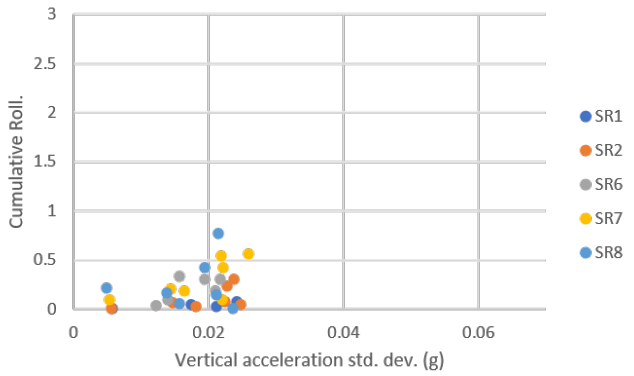


Figure 76. Site A.

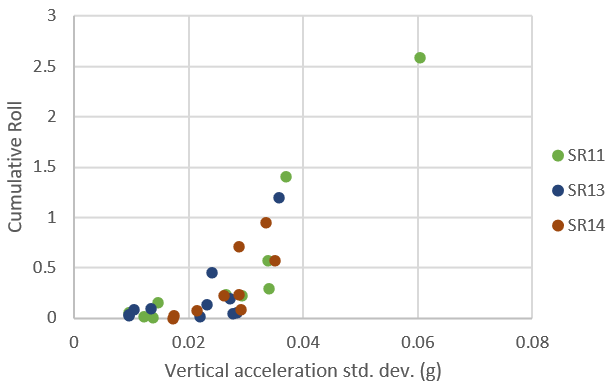


Figure 75. Site B.

This result was further investigated by showing the difference in these summary statistics between selected pairs of SmartRocks, shown in Table 10. These pairs of SmartRocks were selected based on the relative location of SmartRocks, either gauge-side or field-side and either in front of or behind ties with respect to traffic, at both Site A and Site B. Mathematically, this was done by taking the cumulative Roll difference SR14 − SR6 plotted against the vertical acceleration standard deviation difference SR14 − SR6, for example. This is not a perfect measure, since SR8 does not have a functioning SmartRock counterpart at Site B. However, the resulting differences in summary statistics, shown in Figure 77, difference of cumulative Roll and Pitch vs. difference of vertical acceleration standard deviations for selected SmartRock pairs in Table 10, illustrate very clear trends in SmartRock behavior.

Table 10. Table showing selected pairs of SmartRocks.

Pairs:
• SR1 & SR13
• SR2 & SR14
• SR6 & SR14
• SR7 & SR11

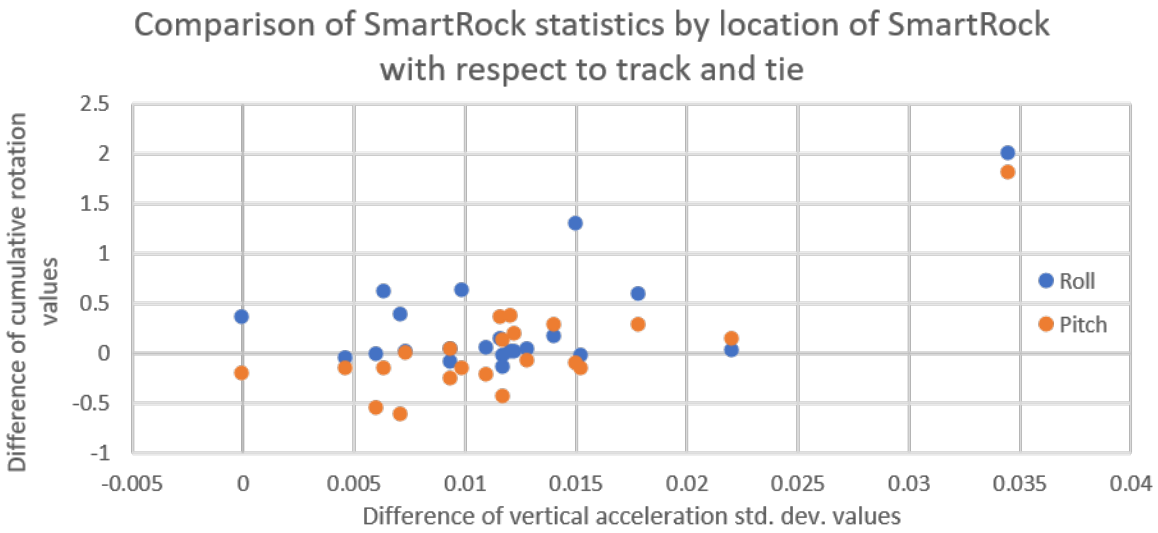


Figure 77. Difference of cumulative Roll and Pitch vs. difference of vertical accel. std. dev. for SmartRock pairs in Table 10.

First, it is again clear that cumulative Roll values for Site B were larger than their counterpart SmartRocks at Site A. This behavior was far more significant when looking at cumulative Roll compared to cumulative pitch. This means that under-tie pads likely have a larger effect on the cumulative Roll of a SmartRock throughout passing trains than on the cumulative Pitch. Furthermore, both the differenced cumulative Roll and Pitch appear to increase linearly with increasing vertical acceleration standard deviation. This makes sense, since it has already been shown, in Figure 74 and Figure 73, that cumulative rotation increases with increasing vertical acceleration standard deviation.

The k-means clustering analyses introduced show a clear difference in SmartRock behavior based on the installation site of the SmartRock (remembering that SR1−SR8 are Site A SmartRocks with no under-tie pads). This was shown in comparisons of lateral and longitudinal acceleration summary statistics and cumulative rotation summary statistics. Furthermore, correlations between SmartRock response differences and their installation location and recording date and subsequent train identification were introduced.

SmartRock response was understood in more specific detail after this analysis. Importantly, it was noted that there is a clear relationship between rotation parameters around the lateral and longitudinal axes and acceleration parameters around the longitudinal and lateral axes, respectively. This is because forces along a particular axis likely torque SmartRocks, causing them to rotate around the perpendicular axis. For example, lateral forces cause changes in Pitch rather than Roll. This is well illustrated by Figure 79 and Figure 78, which show a linear relationship between cumulative Roll and longitudinal acceleration quartile deviations as well and cumulative Pitch and lateral acceleration quartile deviations, respectively. Thus, large vibrations in SmartRocks in the track plane are likely to cause larger changes in the SmartRocks orientation.

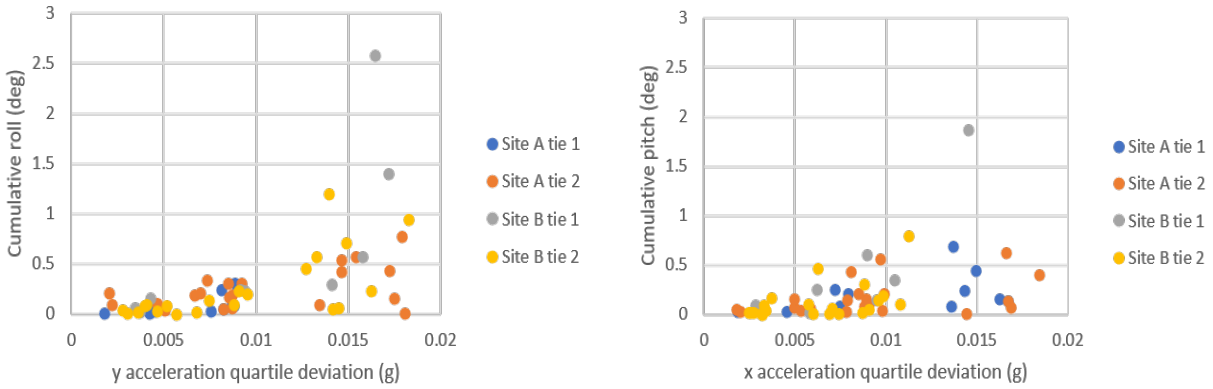


Figure 79. Cumulative Roll vs. longitudinal accel. Figure 78. Cumulative Pitch vs. lateral accel.

LOGISTIC REGRESSION

The k-means clustering results helped inform other analysis techniques. With differences in SmartRock response based on installation site observed, a logistic regression model was created to investigate what SmartRock descriptive statistics indicate an installation site using the generalized linear model *glm()* function in R (R Core Team, 2020). It is expected that this method would be capable of identifying installation sites, since the inclusion of under-tie pads at Site A appears to affect the SmartRock response. Logistic regression is one of the many types of statistical models used to model how independent variables indicate the behavior of (presumed) dependent variables in the dataset. Logistic regression in particular is a binary dependent model that analyzes how the binary dependent output of one variable is correlated with a set of chosen independent variables (Tolles & Meurer, 2016). Logistic regression aims to find the best β parameters that result in the equation below yielding the best possible fit based on available data.

$$y = \begin{cases} 1 & if\ \beta_0 + \beta_1 > 0 \\ 0 & else \end{cases}$$

The dataset for which logistic regression results are introduced is shown in Table 11. This dataset includes SmartRock installation site (*dtsrsite*), vertical acceleration root-mean square (*dtrmsa_z*), cumulative rotation (*dtccrott*), and rotation root-mean-square (*dtrmsrot*) in the track plane (found by taking the vector sum $\sqrt{Roll^2 + Pitch^2}$ of the summary statistics), and the number of vertical acceleration peaks (*dtnumpeaks_z*). These variables were chosen mainly by testing various combinations of variables to see which indicated SmartRock installation site and are summarized in Table 12. As mentioned in the data chapter, since there are only 61 datasets, the number of variables that can be used in regression analysis is limited. General statistical practice is to not have more variables that 10 times the number of samples (Jones et al., 2004).

Table 11. Snippet of dataset used to perform logistic regression.

	dtsrsite	dtrmsa_z	dtccrott	dtrmsrot	dtnumpeaks_z
BNSF_data_test1_acc_20_02_27_14_30	A	0.024187123	0.26002639	0.12058203	355
BNSF_data_test1_acc_20_03_01_14_32	A	0.005832599	0.03066459	0.03286525	252
BNSF_data_test1_mt_1_acc_20_02_24_15_35	A	0.021209129	0.09172439	0.15288042	564
BNSF_data_test1_mt_2_acc_20_02_24_15_35	A	0.017442285	0.25993134	0.13558452	544
BNSF_data_test11_acc_20_02_24_16_00	B	0.038488577	1.44600121	0.58295625	820
BNSF_data_test11_acc_20_02_24_16_40	B	0.029267274	0.23227014	0.47735806	822
BNSF_data_test11_acc_20_02_29_15_30	B	0.026554350	0.34605549	0.42660185	666
BNSF_data_test11_acc_20_03_01_15_48	B	0.033932822	0.83209664	0.55677225	642
BNSF_data_test11_mt_1_acc_20_02_27_15_37	B	0.033972924	0.33314660	0.40167874	700
BNSF_data_test11_mt_1_acc_20_03_02_15_30	B	0.012151346	0.02635997	0.08340177	1202
BNSF_data_test11_mt_1_acc_20_03_09_16_29	B	0.060756884	3.18492337	0.94404371	346
BNSF_data_test11_mt_2_acc_20_02_27_15_37	B	0.009631765	0.11327644	0.07782862	1307
BNSF_data_test11_mt_2_acc_20_03_02_15_30	B	0.014687839	0.17302415	0.16174154	1036
BNSF_data_test11_mt_2_acc_20_03_09_16_29	B	0.013726746	0.01127444	0.03516717	578

Table 12. Summary of logistic regression variables.

Variable	Description
P(SiteB)	Proability sample is a Site B SmartRock per model
dtsrsite	SmartRock installation Site (A=0, B=1)
dtrmsa_z	Vertical acceleration root-mean-square
dtccrott	Cumulative rotation in track plane
dtrmsrot	Rotation root-mean-square in track plane
dtnumpeaks_z	Number of vertical acceleration peaks

This dataset was used to create a logistic regression model with the SmartRock installation location as the dependent variable after coercing it into a binary output (with Site B being 1 and Site A being 0). This logistic regression model yielded the coefficients shown in Table 12 using the model equation below. From this table we see that the coefficients for vertical acceleration root-mean-square (*dtrmsa_z*) and number of vertical acceleration peaks (*dtnumpeaks_z*) variables were determined to be very statistically significant,

while the rotational RMS (*dtrmsrot*) was somewhat significant. Notably, the coefficient values for the acceleration variables were positive, indicating that larger values of these variables indicate a higher likelihood that the two datasets are similar.

$$P(SiteB) = \frac{e^{-10.9+455.4rmsa_z-1.86ccrott-12.45rmsrot+0.008numpeaks\_z}}{1+e^{-10.9+455.4rmsa_z-1.86ccrott-12.45rmsrot+0.008numpeaks\_z}}$$

Table 13. Output of logistic regression.

```
Call:
glm(formula = dtsrsite ~ ., family = "binomial", data = dattrfff)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.10179  -0.57886  -0.05036   0.64594   1.83216

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -10.912917   2.820009  -3.870 0.000109 ***
dtrmsa_z     455.390161 124.384805   3.661 0.000251 ***
dtccrott      -1.863379   1.442683  -1.292 0.196494
dtrmsrot     -12.449267   5.263052  -2.365 0.018010 *
dtnumpeaks_z   0.008149   0.002414   3.376 0.000735 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 84.548  on 60  degrees of freedom
Residual deviance: 49.108  on 56  degrees of freedom
AIC: 59.108

Number of Fisher Scoring iterations: 5
```

This result is significant because it demonstrates the SmartRock's ability to indicate track section conditions. In this case, SmartRock installation site classifications were known and thus a logistic regression could be performed investigating SmartRock site as a binary dependent variable to identify the probability that a track location is of a defined support condition, given the measured SmartRock responses. Figure 80 shows the relationship of SmartRock installation sites to the sum of vertical acceleration root-mean-square and number of peaks based on the logistic regression model. From this plot, as both or either of these values increase, so does the likelihood that the SmartRock dataset was recorded at Site B rather than Site A.

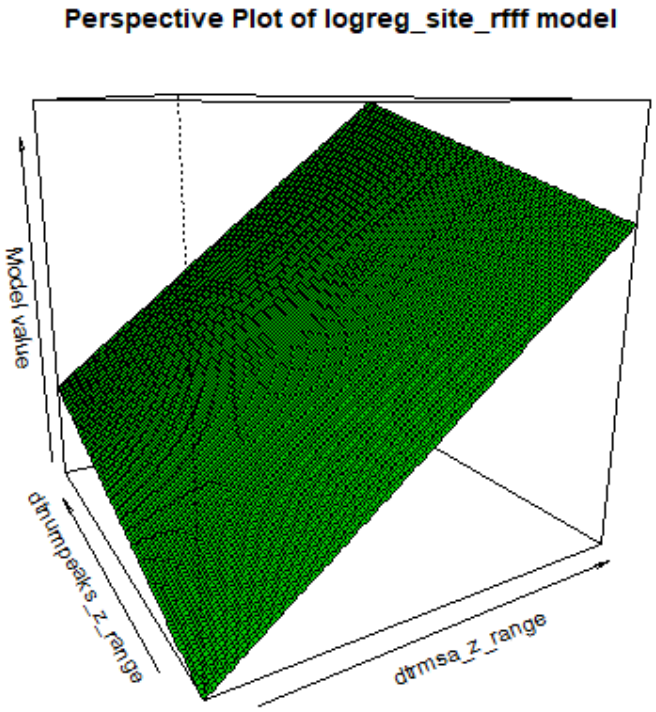


Figure 80. Perspective plots of logistic regression model values (ranging from 0-100% probability that the SmartRock was installed at Site B).

Sensitivity analysis was also conducted on this logistic regression model by calculating a confusion matrix. Samples were determined to be of Site B if the model probability was > 0.5 and of Site A if the model probability was ≤ 0.5. Table 12 shows the confusion matrix of the logistic regression model. A handful of false-positives and false-negatives were present in the model, but with a specificity of 0.8 accuracy, the model performed well. The model also yielded a sensitivity of 0.8, since 24 of the 30 samples predicted to be a part of Site B were actually installed there. Logistic regression models can be analyzed based on the sum of their specificity and sensitivity parameters, which vary from 0 − 1. The calculated parameters for this model showed that it described the data well.

Table 14. Confusion matrix.

	Actual	
Predict	A	B
A	25	6
B	6	24

An extension of this analysis could be used to correlate SmartRock descriptive statistics with ballast condition grades from manual inspection using multinomial logistic regression. This hypothesis and the implication it has on future SmartRock research is discussed in the following chapter.

OTHER RESULTS

This section introduces all other analysis results not previously discussed in this chapter. Many of these analyses were conducted to conclusively investigate methods from which information could be extracted from SmartRock time series. First, the train speeds calculated from the strain gauge data are investigated. Furthermore, Fourier analysis done to put SmartRock response data in the frequency domain is discussed.

Train speeds were calculated from the strain gauge channels and are shown in Table 4. Since train speeds were not available for every SmartRock dataset, these data could not be utilized in the bulk of analysis presented in this chapter thus far. This section covers the use of these train speed values to investigate the correlation between SmartRock response data and strain characteristics measured from the strain gauge rosettes.

As introduced earlier, the time a train took to pass over a given track section was calculated from the datetime data included in the SmartRock datasets. From these values, the length of a train (in feet) could be approximated by multiplying the train speeds estimated from the strain gauge data and the approximated time it took for a train to pass. This incorporates both descriptive statistics in a meaningful way to categorize passing trains.

The approximated train lengths were plotted against statistics calculated from the vertical acceleration time series to investigate if train lengths meaningfully correlated with any SmartRock response characteristics. Figure 82 shows vertical acceleration peak counts plotted against approximated train lengths for available strain gauge datasets. As expected, each train passing has similar approximated train lengths, which vary, since train lengths were determined by the approximated time it took for a train to pass per the subsetting method. Additionally, Figure 81 shows vertical acceleration standard deviations versus approximated train lengths. In both plots, clear variations exist in the vertical acceleration statistics, even for the same train passing. These differences can likely be attributed to SmartRock installation site and location differences, since the train passings are the same. Notably, approximated train lengths correlated with acceleration peak counts, indicating that longer trains likely produce more peaks. Furthermore, SR6 and SR14 showed much larger peak counts than the other SmartRocks for train passings recorded on February 24. These differences became smaller for the remaining recorded response data.

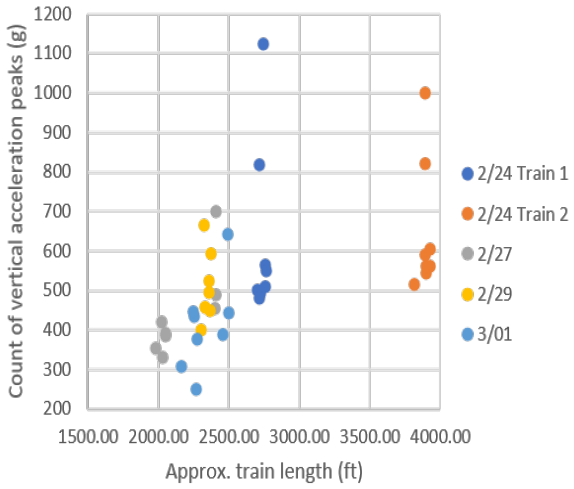


Figure 82. Vertical acceleration peak counts vs. approximated train lengths.

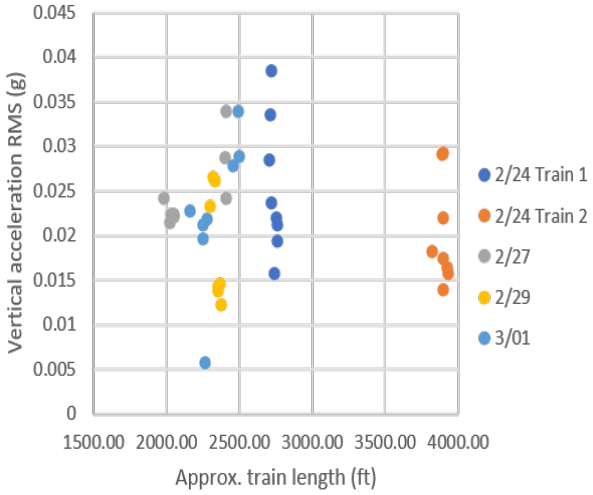


Figure 81. Vertical acceleration root-mean-square vs. approximated train length.

FOURIER ANALYSIS

Fourier analysis was conducted to try and extract the dominant frequency modes present in the SmartRock response data. Fourier analysis is the study of time series as a sum of trigonometric sinusoidal functions after transformation for the time to the frequency domain using a Fourier transform (Kaiser, 1994). Fast Fourier transforms (FFTs) were used to put acceleration and Euler angle time series into the frequency domain. FFTs compute discrete Fourier transforms of a sequence of data by factorizing the discrete Fourier transform matrix into a product of spares factors (Loan, 1992). This greatly reduces computation time, making FFTs one of the most popular methods of performing Fourier transforms.

The discrete Fourier transform is defined by the equation below. Evaluating this definition directly requires a large number of operations, which the fast Fourier transform algorithm reduces without losing accuracy (Loan, 1992). The discrete Fourier transform effectively transforms a finite sequence of equally spaced samples into a sequence of equally spaced samples of the discrete-time Fourier transform, which is a complex-valued function of frequency, of the same length. For the SmartRock response data, the frequency domain can be directly related to the time step of the time series, meaning the frequencies have physical significance.

$$X_k = \sum_{n=1}^{N-1} x_n e^{\frac{-i2\pi kn}{N}} \quad for\ k = 0, \dots, N-1$$

Where

$x_0, \dots, x_{N-1}$ = complex numbers
N = length of the discrete series

For this reason, Fourier analysis on SmartRock response data, and the acceleration time series in particular, can be quite useful. As the plots in the data chapter indicated, the acceleration time series were stationary. Furthermore, the SmartRocks vibrated under dynamic wheel impacts from train passings. Ideally, FFTs would reveal dominant frequency modes that can be correlated with load impulses imposed on the SmartRock. Figure 85 shows an FFT calculated from the vertical acceleration time series shown in Figure

15. The frequency of this FFT is in one over time units (1/seconds) i.e., Hertz. The plot shows a lot of noise with higher amplitudes at lower frequencies. The amplitudes are still notably small, however, mostly since SmartRock acceleration measurements rarely surpassed 0.3 g above or below the mean. From this FFT, it is hard to determine what the dominant frequency modes are. Figure 84 and Figure 83 show similar results calculated for the lateral and longitudinal acceleration series from the same dataset.

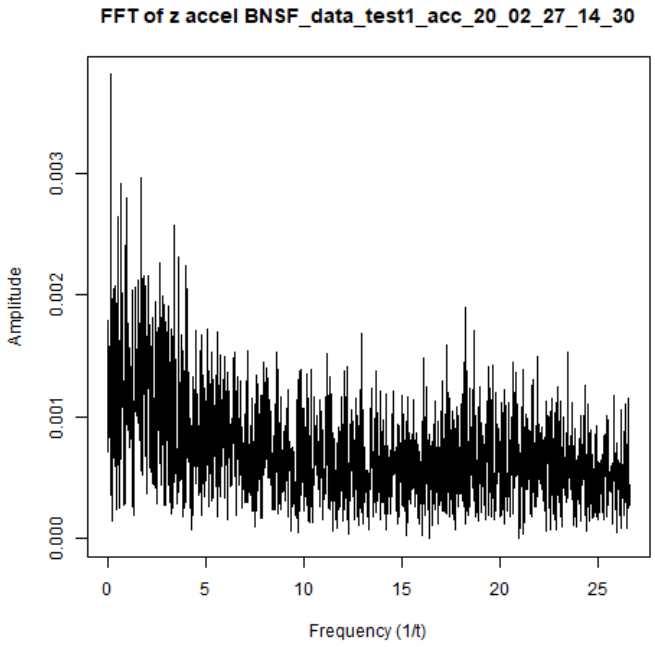


Figure 85. FFT of the vertical acceleration time series shown in Figure 15.

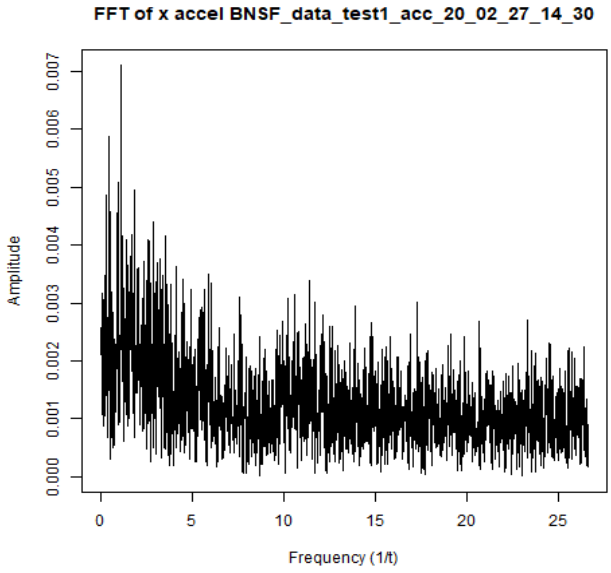


Figure 84. FFT of lateral acceleration.

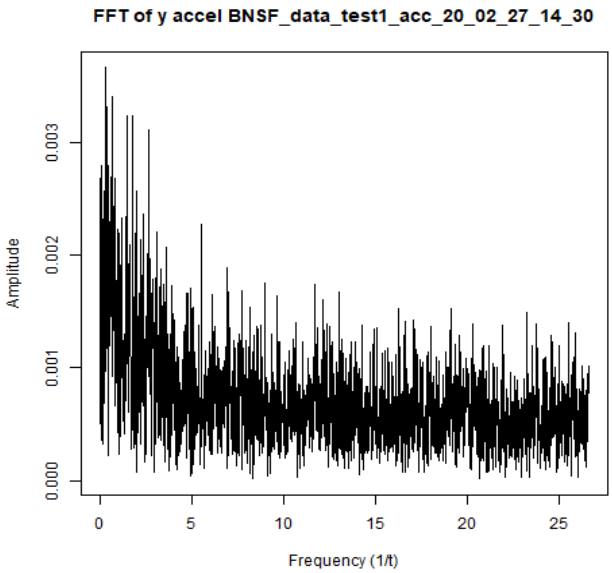


Figure 83. FFT of longitudinal acceleration.

Attempts at noise-removing the time series to reduce the noise in FFT plots were conducted. Application of a moving average filter showed somewhat promising results. Moving average filters smooth sequences of data by replacing data values with the mean of values in a moving window. Since acceleration spikes occurring in the time series typically were only single data points, a small window of $n = 3$ was used such that these acceleration peaks were not obscured. Mathematically, the moving average filter calculates values based on the mean of the values from the original sequence of the data in the window per the equation below, where N is the length of the sequence. The *runMean()* function in R was used to compute the averaged time series for efficiency (R Core Team, 2020).

$$\overline{x_3} = \frac{1}{3}\sum_{1}^{N}\sum_{i=1}^{3} x_i$$

Figure 86 shows both the original acceleration signal and the smoothed signal after a moving average filter of size $n = 3$ was applied. While smoothing did occur, some of the acceleration peaks seen in the raw time series are obscured in the smooth series. Furthermore, dominant frequencies do not present themselves in the FFT plot calculated for the smoothed series shown in Figure 87. The dominant frequencies in this FFT were mostly the same except at higher frequency values, where a pinching effect is occurring due to the moving average filter. Although the acceleration time series was smooth, as desired, this method of noise removal did not help identify dominant frequency modes in the SmartRock response data. Other attempts at noise removal yielded analogous results.

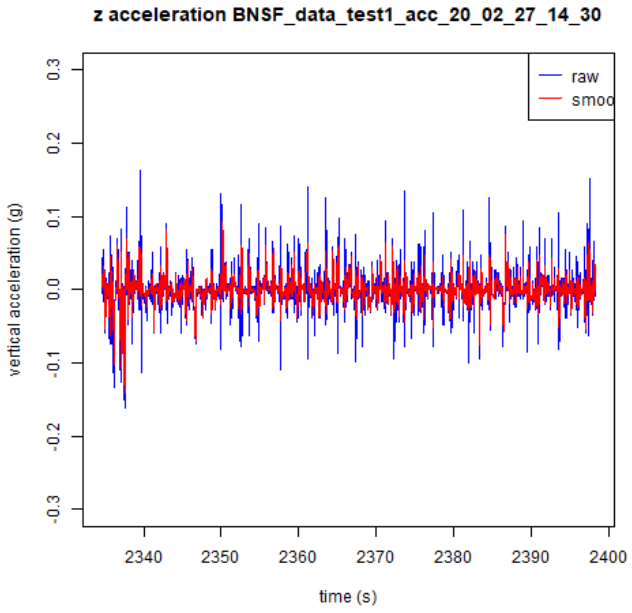


Figure 86. Vertical acceleration after smoothing.

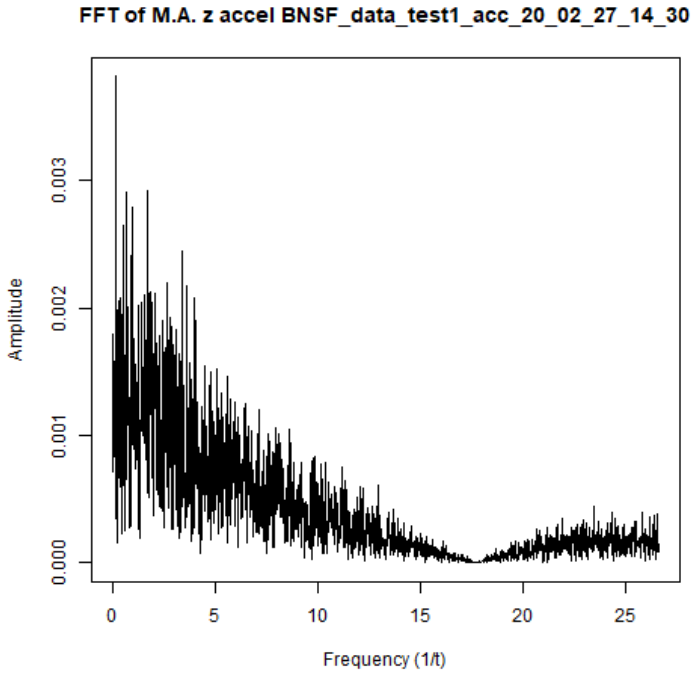


Figure 87. FFT of smoothed vertical acceleration shown in Figure 86.

Fourier analysis was also conducted on Euler angle time series. Since the Euler angle time series were not stationary, Fourier analysis was conducted on the differenced series. Results calculated from the raw rotation data showed negligible amplitude after only a couple of frequency steps. The calculated FFTs for the differenced rotation series are shown in Figure 89 and Figure 88 for the differenced Roll and Pitch, respectively, of the equivalent dataset shown in previous Fourier analysis. Since the differenced Euler angle series were most often relatively flat, except for several instances where larger changes in rotation would occur, the data sequences in the frequency domain were largest at low frequency values. Furthermore, some cyclic behavior was observed in these FFTs, as demonstrated by the rolling peaks observed at frequencies of 3, 6, 8, 10, 12, 15, and 18 Hz. This behavior has not been explained physically. It is important to note that even the most dominant frequency modes yielded small amplitudes on the order of 10^{-4}. Thus, no significant dominant frequency can be extracted from these plots, as expected, after the results observed in the acceleration FFTs.

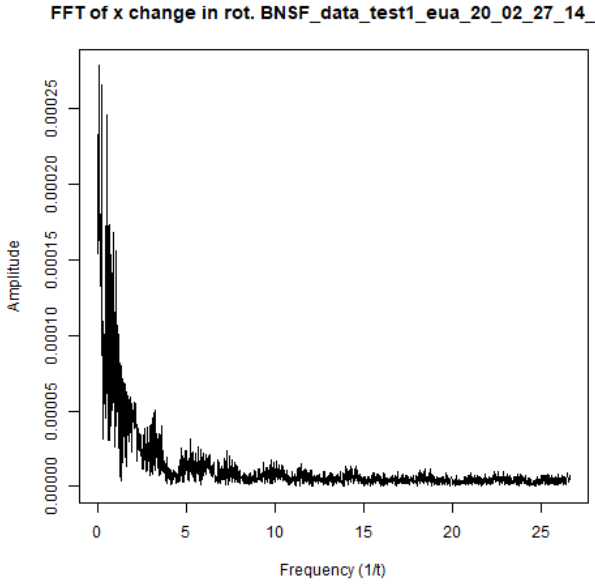


Figure 89. FFT of Roll.

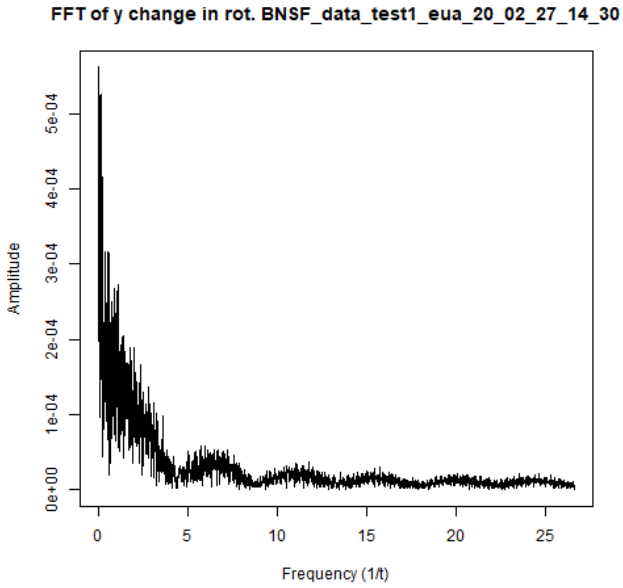


Figure 88. FFT of Pitch.

Although Fourier analysis was not able to identify dominant frequency modes, it showed that SmartRock response data had some cyclic behavior at low frequencies. The frequencies at which peaks occur would be more revealing with additional information about the railway infrastructure and traffic. With just train speeds and SmartRock installation depth, not enough information is available to relate SmartRock dominant frequency modes to load impulses from passing trains. Load impulses felt by track substructure components are felt differently based on the depth of interest but also based on the axle and truck spacing of the railway traffic.

DISCUSSION

Utilizing unsupervised machine learning techniques to analyze SmartRock response data based on descriptive summary statistics allowed for conclusions to be made about what categorical classifiers may be causing differences observed in their response. Furthermore, these techniques allowed for installation differences in the SmartRocks, particularly installation site and the presence/lack of under-tie pads, to be linked to descriptive statistics derived from the response data. These discoveries show that analyzing SmartRock response data using descriptive summary statistics and unsupervised machine learning techniques can yield interesting information about railways and their traffic. Specifically, the analysis showed differences in track support condition and potentially ballast condition, but not enough data were available to definitively make this conclusion.

K-means clustering was applied to investigate the relationship between descriptive statistics calculated from the SmartRock response data. Clustering successfully showed both expected statistical relationships and unexpected ones. In particular, variations in SmartRock acceleration along the tie (laterally) and rail (longitudinally) seemed to depend on the installation location and site of the SmartRock. This means that vibrations due to passing trains were felt differently based on differences in the SmartRocks or SmartRock's installation such as the number of the SmartRock, the location of the SmartRock with respect to ties and rails, and the recorded train passing and the number of the SmartRock.

This chapter also showed the correlation of cumulative Roll values with both SmartRock installation site as well as vertical acceleration standard deviation values. Evidence was provided that shows a clear effect

of the under-tie pads on the cumulative Roll of SmartRocks throughout train passings. Surprisingly, this effect was far more prevalent for cumulative Roll values than for cumulative Pitch values.

Logistic regression was used to quantify the link between vertical acceleration variations and support condition, which was emulated using under-tie pads. Smaller variations in vertical acceleration root-mean-square values for SmartRocks under-tie pads means that they are likely working as intended, dampening dynamic wheel impacts on the ballast layer.

CHAPTER 5

Concluding Remarks

RESEARCH SUMMARY

The focus of this report was to introduce and demonstrate methods of analyzing SmartRock response data. The differences identified between SmartRock datasets were linked to various characteristics of the SmartRocks such as the SmartRock number, installation location, and recorded train passing illustrating that ballast-resembling measurement devices can determine track response associated with existing conditions, and subsequently can be used in diagnosing track issues. Track systems are expensive and require routine maintenance, and all quantitative measures that can diagnose their condition are useful to maintenance engineers.

Much research has been conducted in the laboratory and field to understand track deterioration and how it relates to the degradation of track support layers. Measurements of track geometry parameters, both manually and automatically, have historically been used to investigate track condition. Track geometry typically degrades simultaneously with ballast layers in ballasted track. However, maintenance aimed at correcting track geometry often does not improve ballast characteristics and can negatively impact them. In particular, ballast tamping disturbs the ballast, after which compaction, naturally or mechanically, is required to return the layer to its full strength. Since good track geometry condition is required by the FRA for railroad operations, maintenance practices of track and ballast are continually striving to become more effective and efficient.

Quantitative data on track geometry in three dimensions are readily available and regularly collected for most railroad sections. Conversely, measurements of track substructure characteristics and condition are carried out less often. Non-destructive methods of doing so, such as the cone penetrometer tests or ground penetrating radar measurements, measure the substructure condition statically. SmartRocks are wireless sensors that provide new insight into ballast dynamics, since they are capable of measuring ballast behavior under passing trains.

Research presented in this report explored available SmartRock data from pre-processing to analysis. This included equalizing the coordinate basis of each SmartRock with the pre-defined track coordinate basis, extracting train passings, and calculating descriptive summary statistics and distributions from the recorded train passings. The SmartRocks accurately measured their vibration under train passings such that differences in their responses could be identified and explored.

Unsupervised machine learning techniques, k-means clustering, and logistic regression were used to relate differences seen in SmartRock responses to categorical installation differences between the SmartRocks, particularly the track site at which they were installed, one of which contained tie pads underneath sleepers. The effect of the tie pads was reflected in the SmartRocks vibrational behavior.

RESEARCH IMPLICATIONS

The research presented in this report shows that SmartRocks are a viable system of measuring ballast dynamics, confirming the early results of previous, limited research. From the ballast response data, conclusions were able to be drawn from differences in SmartRock installation parameters and train passings. These determinations were made without knowledge of track geometry parameters, train axle loads, or ballast grades based on manual inspection. Still, SmartRock responses were able to be correlated with known information such as the recording dates of trains and differences in installation parameters such as SmartRock location and track site.

The research presented also shows the capability of SmartRocks to identify differences in track section characteristics using the response data alone. Future installations of SmartRock measurement devices, with proper consultation, could be used to identify differences in track section characteristics more specifically.

CHALLENGES

The data used for this research were incomplete. This was demonstrated by the abundance of unusable datasets that did not properly measure a passing train. Additionally, of the 16 SmartRocks, only half of them functioned properly. While the functioning SmartRocks were able to measure acceleration and orientation properly, they did not measure internal stress as designed.

In addition, numerous characteristics of this dataset were unknown. Information about the dataset was gathered from other parties studying SmartRocks who facilitated their installation. This meant that certain information about the type and frequency of railway traffic at the installation site were not collected and thus were unknown. Furthermore, the effect of other possible track issues such as track geometry defects could not be considered, since this information was not available. Lastly, and most importantly, the condition of the ballast at both installation sites based on field observations was unknown.

This research was intended to cover a new dataset installed by the researching party at local track locations with the primary goal of studying ballast condition. However, due to the ongoing COVID-19 pandemic, this installation was canceled, and readily available data were instead utilized.

CONCLUSIONS

SmartRock response data were explored and discussed in this report. All SmartRock data were in the form of time series measured in local coordinates. To understand what the measurements were identifying, all data had to be converted to the same track coordinate basis. This involved analyzing the SmartRock orientation data, which were measured as quaternions. To find the initial coordinate basis of each SmartRock and perform a basis change to the global coordinate basis, the first measured quaternion had to be integrated. This research involved taking relatively complex data and turning them into series that clearly illustrated train passings, which was achieved successfully.

Exploratory data analysis was utilized to inspect the shape of the SmartRock time series extracted after pre-processing. The pre-processing procedure was verified by showing that SmartRock acceleration series vertical to the track plane had a mean of 1g. It was determined that the acceleration data were stationary by investigating their autocorrelation. From the response data, descriptive statistics such as the five-number summary, quartile deviations, cumulative rotation, and peak count were calculated, which effectively summarized each dataset and were used in further analysis.

Importantly, the data showed explicitly via EDA on summary statistics that SmartRocks installed at the site without under-tie pads measured larger vertical vibration values. This can be assumed to be due to dynamic wheel impacts from passing trains, since SmartRocks at both sites measured the same trains. Unsupervised machine learning applications were able to verify this by linking specific variations in SmartRock responses to train passings and installation differences. This was done by both exploring k-means clustering and logistic regression analysis results.

First, it was shown that lateral and longitudinal acceleration quartile deviations and standard deviations acted differently, as shown in Figure 57 through Figure 67. This analysis showed that numerous differences in the SmartRocks and their installation locations caused differences in response. Next, cumulative Roll was shown to be larger in magnitude with increasing vertical acceleration standard deviations as well as Roll standard deviations and longitudinal acceleration quartile deviations. These relations were exacerbated for Site B SmartRocks in particular. This was shown specifically in Figure 77, where cumulative Roll values were observed to be much larger for Site B SmartRocks compared to their Site A counterparts. Notably, this behavior was not observed as dramatically for cumulative Pitch values, indicating that under-tie pads have a clear impact on mitigating SmartRock's rotations around the lateral (tie) axis.

As noted, this research was based on a very limited dataset. Results presented in this report should be expanded upon using a larger and more conclusive dataset. Extending the techniques herein to a set of SmartRock data with fixed internal stress measurements would yield more information about ballast dynamics and behavior. Furthermore, identifying other parameters regarding the site at which each SmartRock is installed would greatly improve the ability to analyze the results.

Integrating inspection data, specifically ballast grades, with the SmartRock data would make it possible to correlate SmartRock response differences with verifiable and quantifiable ballast condition and is the most critical area requiring further research. This would introduce the possibility of creating a model using advanced machine learning techniques such as a neural network or Bayesian network that can be trained to assess ballast condition based on the SmartRock response. Ballast condition based on field observations, and no information explicitly correlated to it, were not available in this dataset.

Additionally, train passing information should accompany the SmartRock dataset to make the most use out of their responses. Some SmartRock response differences can be attributed to differences in railway traffic. If these differences are quantified independent of using the SmartRock data, variations in railway traffic may be accounted for in the SmartRock response data (e.g., variations can be attributed to substructure differences).

Lastly, installing a larger number of SmartRocks in more diverse track sections would diversify the datasets even more, potentially revealing other areas of response differences. It is suggested that four SmartRocks be installed in each track section to provide multiple datasets that should measure similar results. Beyond this, multiple track sections should be chosen for SmartRock installation, keeping installation parameters equivalent between sites, such that differences in each set of SmartRocks can be attributed to differences in site conditions.

References

Chrismer, S. M., & LoPresti, J. (1996). *Track Substructure Assessment with the Cone Penetrometer*. Transportation Technology Center, Association of American Railroads, Pueblo, Colorado. https://doi.org/TD-96-002

Dunn, P., & Davis, M. P. (2005). *Measurement and Data Analysis for Engineering and Science*. McGraw-Hill.

Everitt, B. (1998). *Dictionary of Statistics*. Cambridge University Press.

Federal Railroad Administration (2008). *Track Safety Standards Compliance Manual*. United States Department of Transportation.

Federal Railroad Administration (2019). *FRA Office of Safety Analysis*. United States Department of Transportation.

Gagniuc, P. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons.

Gentle, J. E. (2007). *Matrix Transformations and Factorizations*. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-64867-5_5

Hamid, A., Rasmussen, K., Baluja, M., & Yang, T.-L. (1983). Analytical Descriptions of Track Geometry Variations. *Transportation Research Record*, *838*:19-26.

Hamilton, W. R. (1843). *On Quaternions; Or A New System of Imaginaries in Algebra*. Letter to John T. Graves.

Hamilton, W. R. (1853). Quaternion quotient lines and tridimensional space time. *Hodges and Smith*, 60.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C*, *28*(1), 100–108. https://doi.org/JSTOR 2346830

Jenkins, D. G., & Quintana-Ascencio, P. F. (2020). A solution to minimum sample sizes for regressions. *PLoS ONE*, *15*(2).

Jones, S. R., Carley, S., & Harrison, M. (2004). An introduction to power and sample size estimation. *Emergency Medicine Journal 20*(5):453-8 .

Kaiser, G. (1994). A Friendly Guide to Wavelets. *Physics Today*, *48*(7), 57–58.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441–458.

Kuhn, M. (2020). *caret: Classification and Regression Training*.

Li, D., Hyslip, J., Sussmann, T., & Chrismer, S. (2016). *Railway Geotechnics*. CRC Press Taylor & Francis Group.

Liu, S., Huang, H., Qiu, T., & Gao, L. (2017). Comparison of Laboratory Testing Using SmartRock and Discrete Element Modeling of Ballast Particle Movement. *Journal of Materials in Civil Engineering*, *29*(3).

Liu, S., Huang, H., Qiu, T., & Kerchof, B. (2019). Characterization of Ballast Particle Movement at Mud Spot. *Journal of Materials in Civil Engineering*, *31*(1).

Loan, C. Van (1992). *Computational Frameworks for the Fast Fourier Transform*. SIAM.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkely Symposium on Mathematical Statistics and Probability*, 281–297.

Milborrow, S. (2020). *earth: Multivariate Adaptive Regression Splines*.

Papoulis, A. (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240–242.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for

Statistical Computing. https://www.r-project.org/
Shenton, M. J. (1978). Deformation of Railway Ballast under Repeated Loading Conditions. In *Railroad Track Mechanics and Technology* (pp. 405–425). Elsevier. https://doi.org/10.1016/b978-0-08-021923-3.50025-5
Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, *5*, 1–34.
Sussmann, T. R., O'Hara, K. R., Kutrubes, D., Heyns, F. J., & Selig, E. T. (2001). Development of ground penetrating radar for railway infrastructure condition detection. In *Proceedings of the Symposium on the Application of Geophysics to Engineering and Environmental Problems*.
Tolles, J., & Meurer, W. J. (2016). Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, *316*(5), 533–534. https://doi.org/10.1001/jama.2016.7653
Wei, T., & Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix*.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the {tidyverse}. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686
Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika Trust JSTOR*, *2334*(448).
Wilson, E. B. (1923). First and second laws of error. *JASA*, *18*(143).
Wuertz, D. (2020). *timeSeries: Financial Time Series Objects (Rmetrics)*.

Yurlov, D., Zarembski, A. M., Attoh-Okine, N, & Palese, J. W. (2019). Probabilistic Approach for Development of Track Geometry Defects as a Function of Ground Penetrating Radar Measurements. *Journal of Transportation Infrastructure Geotechnology*, *6*(1), 1-20, DOI .1007/s40515-018-0066-x

Zarembski, A. M., & Bonaventura, C. (2010). Dynamic Effects of Track Surface Condition on Vertical Wheel/Rail Forces and Energy Consumption. *ASME/IEEE Joint Rail Conference*, *2*. https://doi.org/JRC2010-36002
Zarembski, A. M., Yurlov, D., Palese, J., & Attoh-Okine, N. (2020). *Relationship Between Track Geometry Defects and Measured Track Subsurface Condition*. Final Report submitted to United States Department of Transportation, Federal Railroad Administration.

Zarembski, A. M., Yurlov, D., Palese J. W., Attoh-Okine N, & Thompson, H. (2018). Relationship between Track Geometry Degradation and Subsurface Condition as Measured by GPR, *American Railway Engineering Association Annual Conference*, Chicago, IL, September 2018.

Zeng, K., Qiu, T., Bian, X., Xiao, M., & Huang, H. (2019). Identification of ballast condition using SmartRock and pattern recognition. *Construction and Building Materials*, *221*, 50–59. https://doi.org/10.1016/j.conbuildmat.2019.06.049