1. Report No. RailTEAM UD-6	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Railroad Infrastructure Health Assessment Using Multiway Data Factorization - A Case for Railroad Track Geometry Data		5. Report Date September 2017 6. Performing Organization Code:
7. Author(s) Offei Adarkwa and Nii Attoh-Okine https://orcid.org/0000-0001-5328-5538		8. Performing Organization Report No. UD-6
9. Performing Organization Name and Address Department of Civil & Environmental Engineering University of Delaware 301 DuPont Hall Newark, DE 19716		10. Work Unit No. 11. Contract or Grant No. 69A3551747132
12. Sponsoring Agency Name and Address Office of Research, Development and Technology (RD&T) US Department of Transportation 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

The state of the nation's infrastructure has been of significant interest to the media, policy makers and public. The government is therefore seeking new ways to maximize each dollar spent investing in infrastructure. It is anticipated that there will be a growing demand for railroad infrastructure since federal forecasts have projected a 40% increase in US freight shipments by 2040. To meet this demand, sustained funding must be paired with sound asset management practices. Large amounts of data are generated by both passenger and freight railroad systems in the U.S. and results from the analysis of this data could serve as the basis for proactive maintenance to improve safety and system performance. Different methods have been used to analyze track geometry data but this work focuses on how multiway data analysis can be used to generate insights from this data. The results obtained from this analysis are compared to the two dimensional approach for analyzing the same data in order to showcase the main advantages associated with using multidimensional data analysis techniques in the management of railroads.

17. Key Words Tensor, Multiway Data, Track Geometry	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161. http://www.ntis.gov		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 49	22. Price

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

USDOT Tier 1
University Transportation Center
on Improving Rail Transportation
Infrastructure Sustainability and Durability

Final Report UD-6

RAILROAD INFRASTRUCTURE HEALTH ASSESSMENT USING MULTIWAY DATA FACTORIZATION - A CASE FOR RAILROAD TRACK GEOMETRY DATA

By

Offei Adarkwa, Ph.D.
Department of Civil Engineering
University of Delaware

and

Nii Attoh-Okine, Ph.D.
Department of Civil Engineering
University of Delaware
okine@udel.edu

September 2017

Grant Number: 69A3551747132

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

ABSTRACT

The state of the nation's infrastructure has been of significant interest to the media, policy makers and public. The government is therefore seeking new ways to maximize each dollar spent investing in infrastructure. It is anticipated that there will be a growing demand for railroad infrastructure since federal forecasts have projected a 40% increase in US freight shipments by 2040. To meet this demand, sustained funding must be paired with sound asset management practices. Large amounts of data are generated by both passenger and freight railroad systems in the U.S. and results from the analysis of this data could serve as the basis for proactive maintenance to improve safety and system performance. Different methods have been used to analyze track geometry data but this work focuses on how multiway data analysis can be used to generate insights from this data. The results obtained from this analysis are compared to the two dimensional approach for analyzing the same data in order to showcase the main advantages associated with using multidimensional data analysis techniques in the management of railroads.

Keywords: Tensor, Multiway Data, Track Geometry.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The state of the nation's infrastructure has been of significant interest to the media, policy makers and public hence, the government is seeking new ways to maximize each dollar spent investing in infrastructure. On account of this, there is an increasing need to manage current assets to ensure they function at desired operating levels. There will be a growing demand for railroad infrastructure since federal forecasts have projected a 40% increase in US freight shipments by 2040 and to meet this demand, sustained funding must be paired with sound asset management practices.

Large amounts of track geometry, rail defect, traffic and tonnage, and vertical track interaction (VTI) data generated by both passenger and freight railroad systems in U.S. serve as the basis for proactive maintenance to improve safety and system performance. Over the years, several analysis methods have been used to gain useful insights from railroad data. This work focuses on track geometry data. It is important to explore and analyze this data because poor track geometry can increase the risk of derailment as well as damage to bogie and wheels. Track geometry parameters that are measured during track monitoring include Gage, Crosslevel, Elevation, Warp and Cant. Rail track geometry is usually measured using track recording coaches (TRC) or unattended measurement systems mounted on in-service vehicles.

Multiway data is ubiquitous and the field of railroad engineering is no exception. With large amounts of data being collected at several locations on the railroad track over time, a multidimensional approach to analyzing data may be beneficial. In effect, this work focuses on how multiway data analysis can be used to generate insights from railroad track geometry data. The main objective of this report is to highlight the potential benefits of multidimensional data analysis techniques in the management of railroad infrastructure. Multiway data analysis is an extension of two-way data analysis to higher-order data sets. A multidimensional structure for the data set is justified for track geometry data set since the slices for each inspection date (time step) can be considered as correlated. There are two main decomposition or factorization models for multidimensional data. They are Tucker model and PARAFAC model explained below in the case of a three-way data array. PARAFAC decomposition which is a simpler model to fit compared to the Tucker Decomposition was used to analyze the data set.

Data used for the analysis spanned a 1-mile long section of track with observations recorded at 28 inspection dates (time instances) between June 2013 and April 2016. The track under consideration is a tangent section with a shallow curve on one end. Track geometry variables included in the data set were Gage, Crosslevel, Surface, Alignment and Warp. Descriptive statistics were produced for the quantitative variables in the dataset at the 5268 locations for all 28 observation periods. The mean for surface and alignment measures on both right and left tracks were 0in. The median values for these variables were also close to zero indicating a track section in a good condition. The maximum gage width was 57.325in, which is greater than the gage width threshold of 57.25in for this class of railroad. The surface measurements had the highest standard deviation values. The minimum and maximum alignment values indicated more extreme measurements for the left track compared to the left track.

Track geometry data was centered and scaled before PARAFAC decomposition was carried out on the track geometry data. After 18 iterations, the 3-component model was chosen for further analysis because it explained almost half of the systematic variation in the data (45%) with a high CORCONDIAG value. To ensure whether the right number of components were extracted, split-half analysis is used to validate the model. With the split-half analysis yielding similar results to the decomposition of the entire data set, the 3-component model is suitable to explain trilinear variation in the data. The 3-way model for the track geometry data was compared with a two-way model generated by the principal component analysis (PCA) to identify the benefits of a multiway analysis approach. For each year, there was a consistently high correlation between the surface levels on right and left rail tracks, which the PCA approach failed to capture. The three-way approach was able to capture a more accurate temporal signature of the data set compared with the 2-way approach.

The multiway decomposition approach revealed Surface and Gage measurements as being the most dominant variables responsible for almost half of the variation in the data set. Two distinct groups for inspection dates were also revealed by multiway analysis. Right and left surface measurements were shown to be the most highly correlated pair implying that only one of these can be used in further modeling of the data.

INTRODUCTION

The state of the nation's infrastructure has been of significant interest to the media, policy makers and public. With its limited resources, the government is seeking new ways to maximize each dollar spent investing in infrastructure. In effect, there is an increasing need to manage current assets to ensure they function at desired operating levels. There will be a growing demand for railroad infrastructure since federal forecasts have projected a 40% increase in US freight shipments by 2040 (ASCE, 2017). To meet this demand, sustained funding must be paired with sound asset management practices.

Large amounts of data, generated by both passenger and freight railroad systems in the country serve as the basis for proactive maintenance to improve safety and system performance. Data collected from railroads can be classified as: track geometry data, rail defect data, traffic and tonnage data and vertical track interaction data (VTI). Over the years, several analysis methods have been used to gain useful insights from railroad data. This work focuses on track geometry data. Track geometry parameters that are measured during track monitoring include gage, crosslevel, elevation, warp and cant (Zarembski, 2011). Rail track geometry is usually measured using track recording coaches (TRC) or unattended measurement systems mounted on in-service vehicles (Weston et al., 2015).

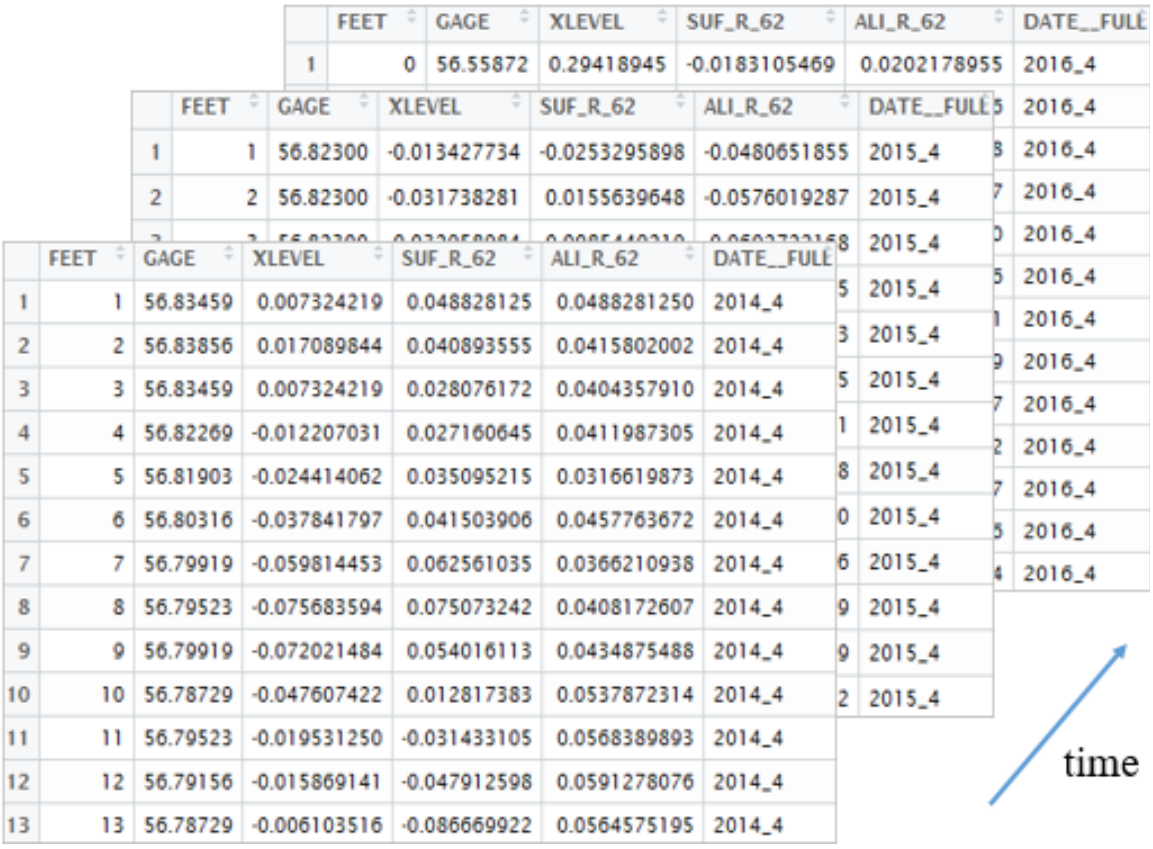
Data from railroad track monitoring can be used to forecast track defect development, verify quality of repairs and improve maintenance management (Nielsen et al., 2013). Track geometry data can be used as input in asset management software for planning maintenance (Lewis, 2011). Also, track geometry data is used for threshold analysis where data points which exceed predefined values are identified for further studies and remedial action. Poor track geometry can increase the risk of derailment as well as damage to bogie and wheels (Lewis, 2011).
Multiway data is ubiquitous and the field of railroad engineering is no exception. With large amounts of data being collected at several locations on the railroad track over time, a multidimensional approach to analyzing data may be beneficial. In the absence of multiway analysis, data is typically coerced into a two-way structure which may sometimes lead to unacceptable simplification of the inherent variation in data (Kroonenberg, 2008). This work focuses on how multiway data analysis can be used to generate insights from railroad track geometry data.

OBJECTIVE

The main objective of this report is to highlight the potential benefits of multidimensional data analysis techniques in the management of railroad infrastructure. Railroad data is inherently multidimensional. The conventional approach to analyzing railroad data views data as having a two-dimensional structure (matrices): rows representing observations with variables in columns. This report highlights the benefits of considering railroad data in a multidimensional sense illustrated in Figure 1 below.

	FEET	GAGE	XLEVEL	SUF_R_62	ALI_R_62	DATE__FULL
1	0	56.55872	0.29418945	-0.0183105469	0.0202178955	2016_4
2	1	56.56268	0.29052734	-0.0183105469	0.0144958496	2016_4
3	2	56.56818	0.28442383	-0.0222778320	0.0072479248	2016_4
4	3	56.57489	0.27832031	-0.0283813477	-0.0003814697	2016_4
5	4	56.58160	0.27465820	-0.0363159180	-0.0083923340	2016_4
6	5	56.58649	0.27221680	-0.0476074219	-0.0144958496	2016_4
7	6	56.58923	0.27099609	-0.0598144531	-0.0194549561	2016_4
8	7	56.59198	0.27343750	-0.0686645508	-0.0225067139	2016_4
9	8	56.59351	0.27587891	-0.0765991211	-0.0255584717	2016_4
10	9	56.59473	0.27709961	-0.0823974609	-0.0289916992	2016_4
11	10	56.59656	0.27465820	-0.0860595703	-0.0339508057	2016_4
12	11	56.59869	0.27221680	-0.0885009766	-0.0396728516	2016_4
13	12	56.60114	0.27099609	-0.0906372070	-0.0469207764	2016_4

(i)



(ii)

FIGURE 1 A sample of railroad defects data (i) matrix-structure (ii) multidimensional structure.

While analyzing railroad data in two dimensions may be the simplified approach, individual differences between observations as well as hidden information on temporal variation may be lost through processes used in simplifying analysis such as averaging. Considering the multidimensional data set shown in Figure 1 (ii), multiway data analysis can help determine the relationship between the track geometry variables with respect to time and across measurement locations simultaneously.

MULTIWAY DATA FACTORIZATION

A multiway array or tensor refers to generalizations of vectors (first-order tensor) and matrices (second-order tensor) (Morup, 2011). An array with an order greater than can be expressed as:

$$\underline{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N} \tag{1}$$

Where $\underline{X}$ is an N[th]-order tensor with dimensions $I_1, I_2, \ldots, I_N$.
For this report, the multiway data structure analyzed is:

$$\underline{T} \in \mathbb{R}^{Location \times Geometry\ parameters \times Inspection\ Dates} \tag{2}$$

Where $\underline{T}$ represents the multidimensional track geometry data shown in Figure 1(ii).

Multiway data analysis is an extension of two-way data analysis to higher-order data sets (Acar and Yener, 2009). In many applications such as environmental data analysis (Stanimirova et al., 2004; Singh et al., 2006), batch process monitoring (Meng et al., 2003), social network analysis (Bader et al., 2008), web link analysis (Kolda et al., 2005) and facial recognition (Vasilescu and Terzopoulos, 2002), analyzing data as two-way arrays limits the level of insight that can be drawn from them. In three-way arrays, there are two main types of subarrays, formed by fixing specific modes in the array. They are fibers and slices. A fiber is formed when two modes in a three-way array are fixed with the remaining mode allowed to vary. On the other hand, a slice is formed when one mode of the three-way array is fixed and the remaining two modes are allowed to vary. See Figure 2 below.

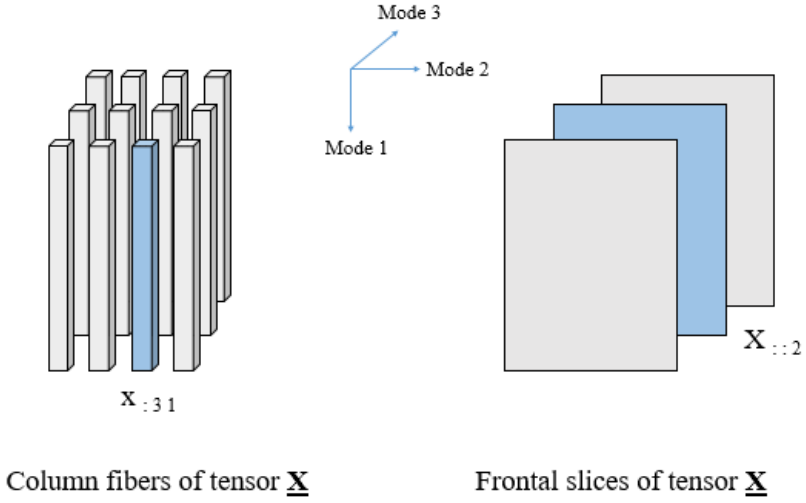


FIGURE 2 Sub-arrays of multiway data.

A multidimensional structure for the data set is justified for track geometry data set since the slices for each inspection date (time step) can be considered as correlated. In other words, inspection data for time *t+1* is dependent on the data for time *t*.

Multiway data is factorized to analyze variation patterns in data. There are two main decomposition or factorization models for multidimensional data. They are Tucker model and PARAFAC model explained below in the case of a three-way data array. The Tucker model for tensor $\underline{X} \in \mathbb{R}^{I \times J \times K}$ decomposes the data such that:

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_{ip} b_{jq} c_{kr} + e_{ijk} \quad (3)$$

Where a_{ip}, b_{jq} and c_{kr} are elements of $A \in \mathbb{R}^{I \times P}$, $B \in \mathbb{R}^{J \times Q}$ and $C \in \mathbb{R}^{K \times R}$ which are loading matrices for modes 1, 2 and 3 respectively. g_{pqr} and e_{ijk} are elements of $\underline{G} \in \mathbb{R}^{P \times Q \times R}$ and $\underline{E} \in \mathbb{R}^{I \times J \times K}$; the core array and residuals array respectively (Acar and Yenner, 2009). The elements of the core array accounts for all possible linear combinations of multiway data (Morup, 2011). The magnitudes of these elements indicate the strength of the relationship between the interacting modes from the loading matrices. See Figure 3 below for illustration of Tucker model (Adapted from Acar and Yenner, 2009).

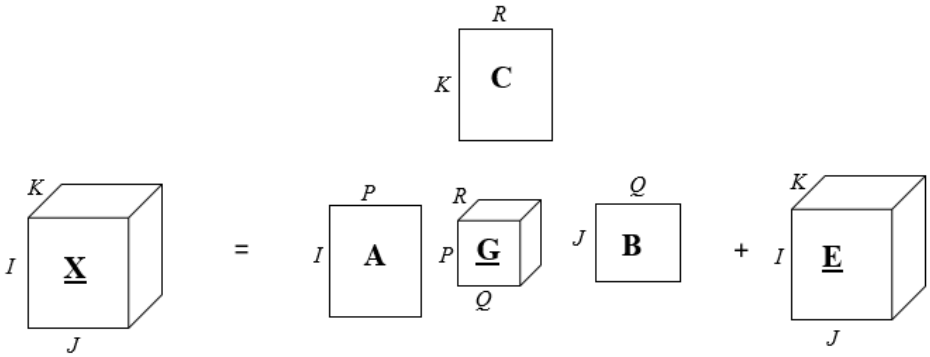


FIGURE 3 Tucker model.

The PARAFAC model can be considered as a constrained Tucker model. In the PARAFAC model, the size of the modes in the core array are identical; *P=Q=R*. Additionally, all the elements in the core array are zero with the exception of those in the leading diagonal. In effect, this leads to interactions between components of the same order making the PARAFAC model a simpler model to estimate. The PARAFAC model can be expressed as:

$$\underline{\boldsymbol{X}} = \sum_{r=1}^{R} a_{ir} \; b_{jr} \; c_{kr} + e_{ijk} \quad (4)$$

Where a_{ir}, b_{jr} and c_{kr} are elements of $A \in \mathbb{R}^{I \times R}$, $B \in \mathbb{R}^{J \times R}$ and $C \in \mathbb{R}^{K \times R}$ which are loading matrices in the first, second and third modes respectively. R is the number of components or columns in each loading matrix. Figure 4 is an illustration of the PARAFAC model.

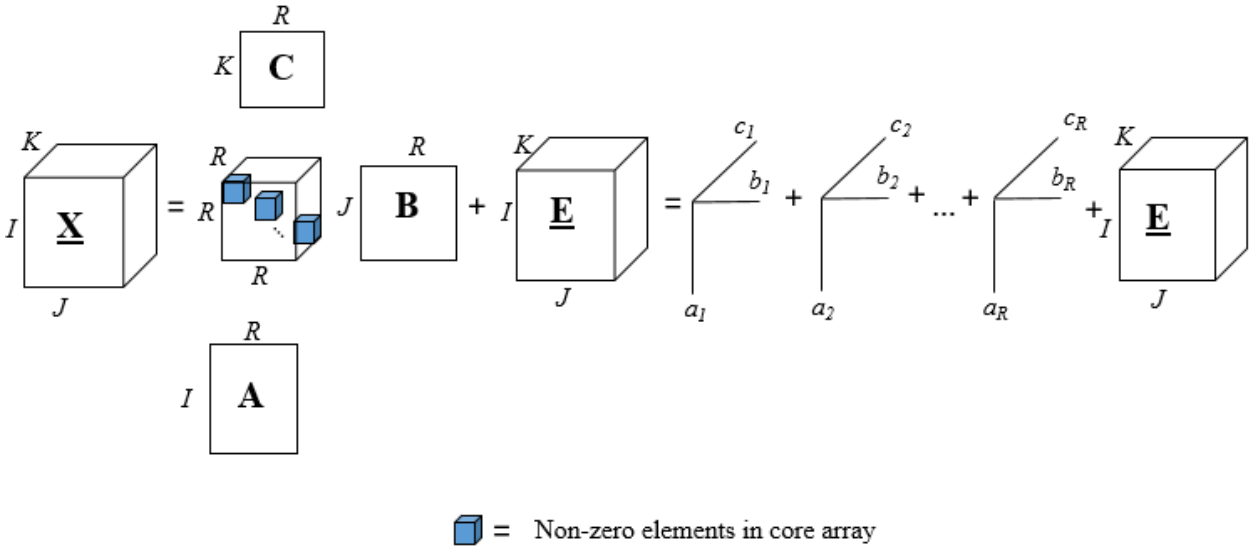


FIGURE 4 PARAFAC model.

The loading matrices obtained after decomposition are used as a basis for describing the data in a condensed form (Bro, 1997), understanding the data and drawing useful insights for further applications.

DATA

Data used for the analysis spanned a 1-mile long section of track with observations recorded at 28 inspection dates (time instances) between June 2013 and April 2016. The track under consideration is a tangent section with a shallow curve on one end. Track geometry variables included in the data set are as follows:

1) Gage: This refers to the distance between two railheads measured at right angles to the rails in a plane 0.625in below the top surface of railheads (UFC, 2008). Track gage measurements, labeled as GAGE were recorded in inches.
2) Crosslevel: The rail crosslevel is the difference in elevation between the top surfaces of two rails measured perpendicular to the tracks (UFC, 2008, FRA 2002). Crosslevel in this data set was measured in inches and represneted by the label XLEVEL.
3) Surface: Surface refers to the relative elevation of two rails along the track. For this data set, surface was measured at the mid-point of a 62-foot chord on the right (SUF_R_62) and left rail (SUF_L_62) tracks in inches.
4) Alignment: Alignment is the perpendicular distance measured in inches at the mid-point of a 62-foot string line stretched on the gage side of right (ALI_R_62) or left rail (ALI_L_62) track at a distance of 0.625 inches below top of railhead (UFC, 2008).
5) Alignment (L-62): This refers to alignment for the left rail track.
6) Warp: This is the difference between crosslevels at two points less than or equal to 62 feet apart (UFC, 2008). It was labeled WARP_62 in the data set.
7) The remaining variables captured location and dates for observations
    a. Location: This refers to the specific location where track variable measurements were taken. It was labeled in the data set as FEET.

b. Inspection Date:_Dates were recorded in the format: year_month. The label for this variable was DATE_FULL.

DATA PREPROCESSING

Data was preprocessed to ensure uniformity in the data before exploratory data analysis. The first preprocessing step involved removing rows at the end of data sets for specific periods to ensure an equal number of observations for each time stamp within the analysis period. Table 1 shows the number of observations for all analysis periods.

TABLE 1 Number Of Observations for Analysis Periods

Data collection period	No. of observations
12/2015	5276
3/2016	5271
4/2016	5272
1/2016, 10/2013, 12/2013, 6/2013, 7/2013, 8/2013, 9/2013	5270
1/2014, 3/2014, 2/2016, 8/2015, 4/2014, 6/2014, 7/2014, 10/2014, 11/2014, 12/2014, 1/2015, 2/2015, 3/2015, 5/2015, 6/2015, 7/2015, 11/2015	5269
4/2015	5268

The number of observations were reduced to 5268 for each inspection date with the exception of April 2015 (4/2015).

EXPLORATORY DATA ANALYSIS

Table 2 shows the descriptive statistics for the quantitative variables in the dataset at the 5268 locations for all 28 observation periods. The mean for surface and alignment measures on both right and left tracks were 0in. The median values for these variables were also close to zero indicating a track section in a good condition. The maximum gage width was 57.325in, which is greater than the gage width threshold of 57.25in for this class of railroad (Track Compliance Manual, 2002). The ideal gage width for railroads is 56.5in, which was 0.1in less than the mean gage width for the track section.

The surface measurements had the highest standard deviation values. From Table 2, the minimum and maximum alignment values indicated more extreme measurements for the left track compared to the left track.

TABLE 2 Descriptive Statistics for Quantitative Variables in Railroad Dataset (6/2013-4/2016)

Variable	Gage	Crosslevel	Surface (Right)	Surface (Left)	Alignment (Right)	Alignment (Left)	Warp
Min.	56.276	-0.630	-0.970	-1.289	-0.794	-0.275	-0.665
1st Qu.	56.638	-0.045	-0.039	-0.037	-0.029	-0.027	-0.051
Median	56.697	0.051	0.002	0.001	0.000	0.000	0.000
Mean	56.676	0.105	0.000	0.000	0.000	0.000	-0.002
3rd Qu.	56.733	0.181	0.044	0.041	0.029	0.027	0.051
Max	57.325	1.234	0.822	0.935	0.410	0.286	0.681
St. Dev.	0.085	0.244	0.108	0.111	0.050	0.046	0.107
Range	1.049	1.864	1.792	2.224	1.204	0.561	1.346
No. of Observations	147504	147504	147504	147504	147504	147504	147504

Gage

Figure 5shows the variation of gage widths at the 5268 track locations in the dataset over the analysis period. It is evident that maintenance works were carried out after August 2015 due to the significant drop in mean gage widths (dots in Figure 5) after this period. Another interesting observation is the gradual rise in mean gage width from January 2016. In April 2016, there were observations with values greater than 57.25in which raises safety concerns for this class of railroad (Track Compliance Manual, 2002). The dots refer to the mean gage width for each analysis period. The cumulative gage width distribution for the three analysis years (Figure 6) showed 99.99% of all the observations were less than or equal to the 57.25in threshold value. This indicates that very few locations had gage widths which were of concern between 2013 and 2016.

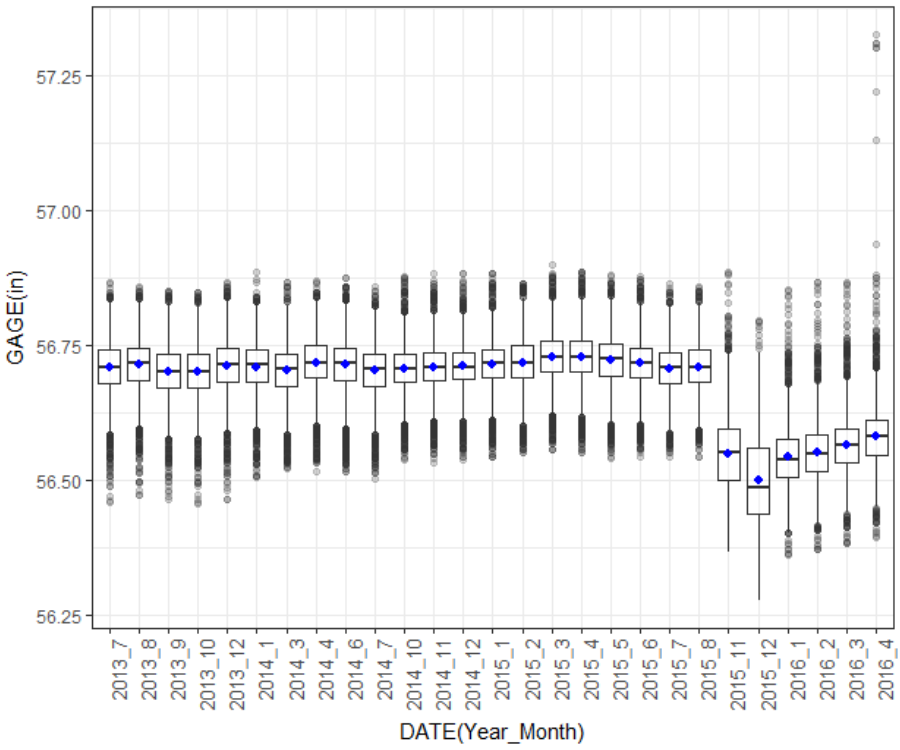


FIGURE 5 Gage width distribution (07/2013-04/2016).

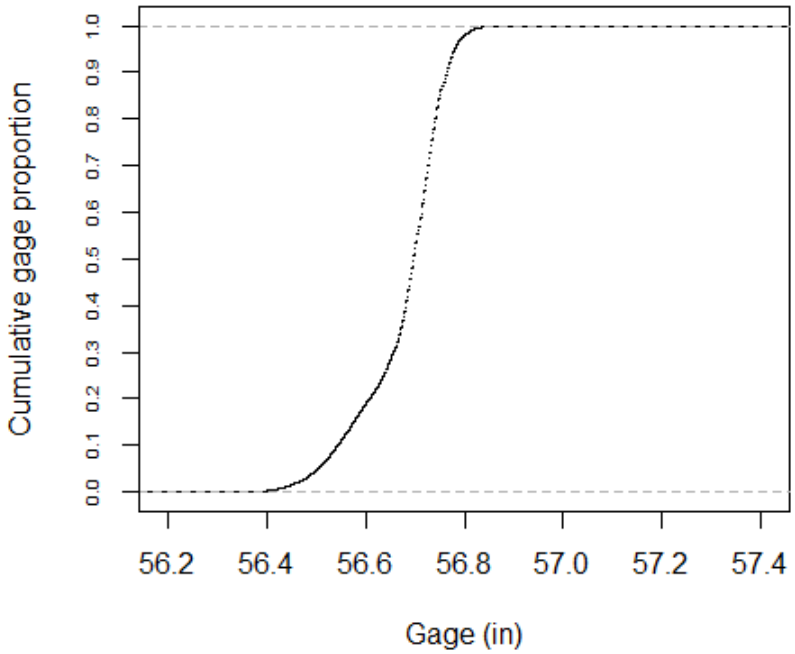


FIGURE 6 Cumulative frequency distribution of gage widths (7/2013-4/2016).

Standard deviations for gage widths at each track location revealed a different pattern. Figure 7 shows the standard deviation for gage widths at all track locations from June 2013 to April 2016. The highest standard deviations for gage widths were recorded at locations between 1400ft-1600ft

and 1800ft-2000ft. This may signal high changes in gage widths from July 2013 to April 2016. The high changes were due to two main reasons: 1) gage widening and 2) maintenance.

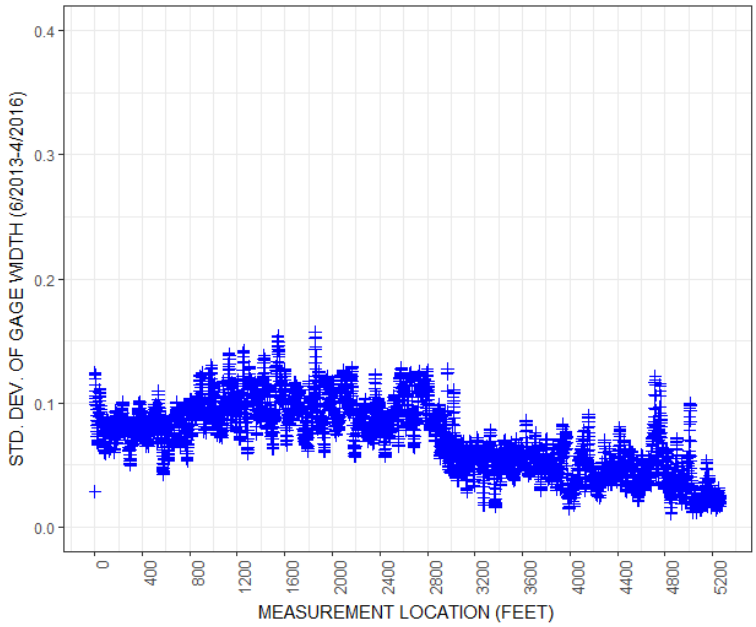


FIGURE 7 Standard deviation of gage width measurements for each location from 06/2013-04/2016.

For each location, the maximum deviation from the ideal gage width from 2013 to 2016 were analyzed. Maximum deviation is expressed as:

$$MaxDgage_i = \frac{\max(gage_i) - 56.5}{56.5} \times 100 \quad (5)$$

Where $MaxDgage_i$ is the maximum gage deviation expressed as a percentage for section i for all the analysis periods and $max\ (gage_i)$ is the maximum gage width for the i^{th} section over the analysis period. Figure 8 shows a continuous section of track between 2800ft and 3000ft experienced high deviations from the ideal gage width of 56.5in. This analysis was done in order to visualize sections which experienced the highest amount of gage widening during the entire analysis period. The continuous section with gage widths greater than 57.25 inches in April 2016 are shown below in Table 3.

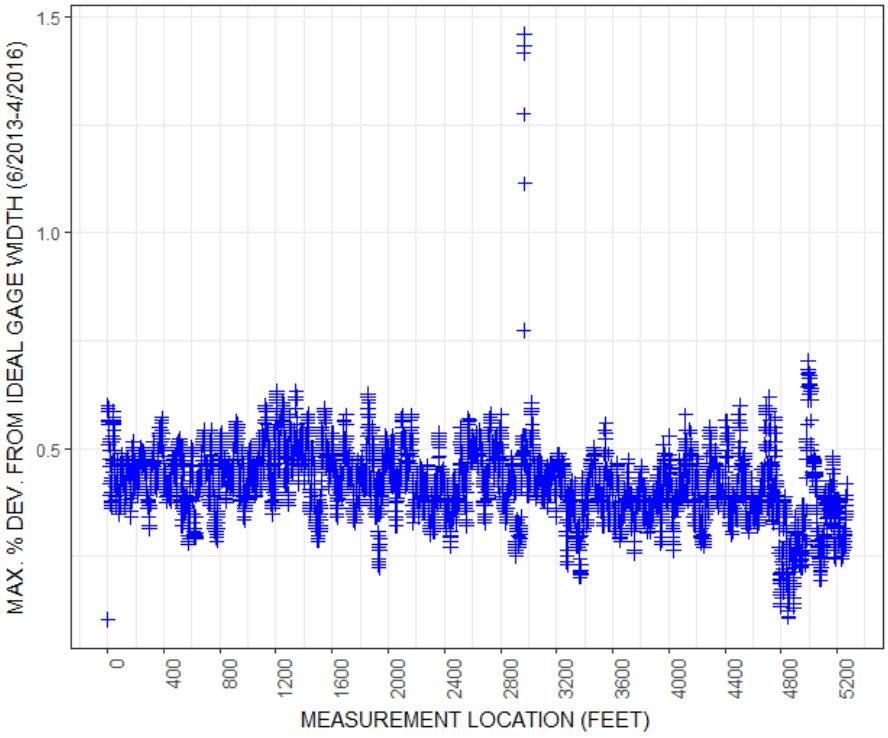


FIGURE 8 Maximum gage width deviation for sections (June 2013- April 2016).

TABLE 3 Sections with Gage>57.25 Inches (April 2016)

Section location	Gage (in)
2967	57.300
2968	57.300
2969	57.325
2970	57.310
2971	57.310

Figure 9 also shows how gage widths changed over the years. It is clear that in 2016, a higher proportion of the observations were below the threshold gage width due to maintenance activities carried out in the previous year. The year 2015, had the most spread out observations which were the result of maintenance works leading to a reduction in gage widths and the corresponding spread of the data.

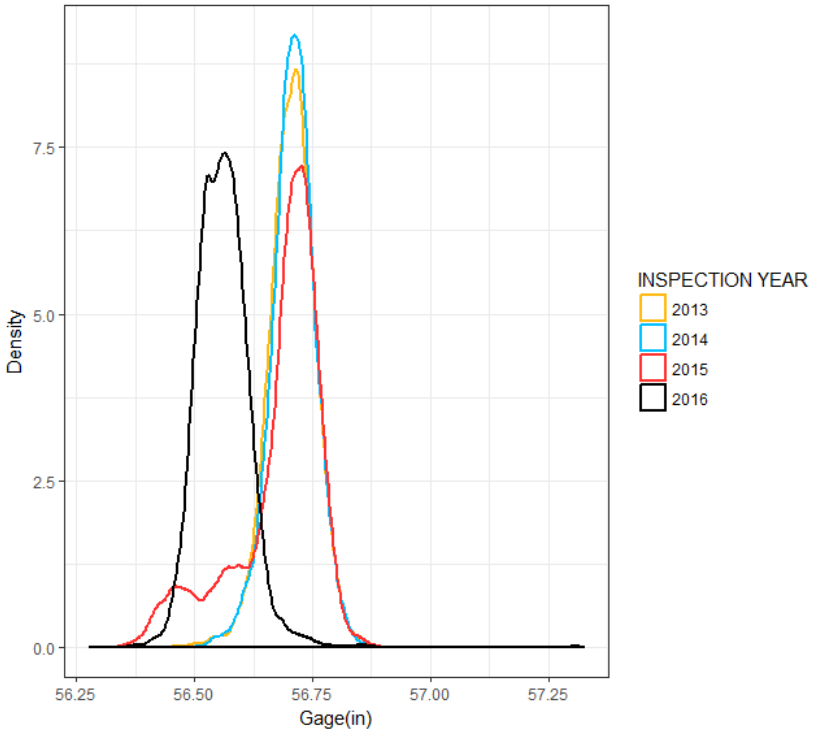


FIGURE 9 Kernel density curves for gage with respect to inspection year.

Crosslevel

Crosslevel is considered as part of environmental variables which contribute to track irregularity (Chaolong et al., 2002). From Figure 10, all observed crosslevel measurements were between +1.5in and -1.0in. Relatively high levels of mean crosslevel values are observed from December 2015 to April 2016. The overall trend for this period was a gradual rise in mean crosslevel as well as maximum positive crosslevel.

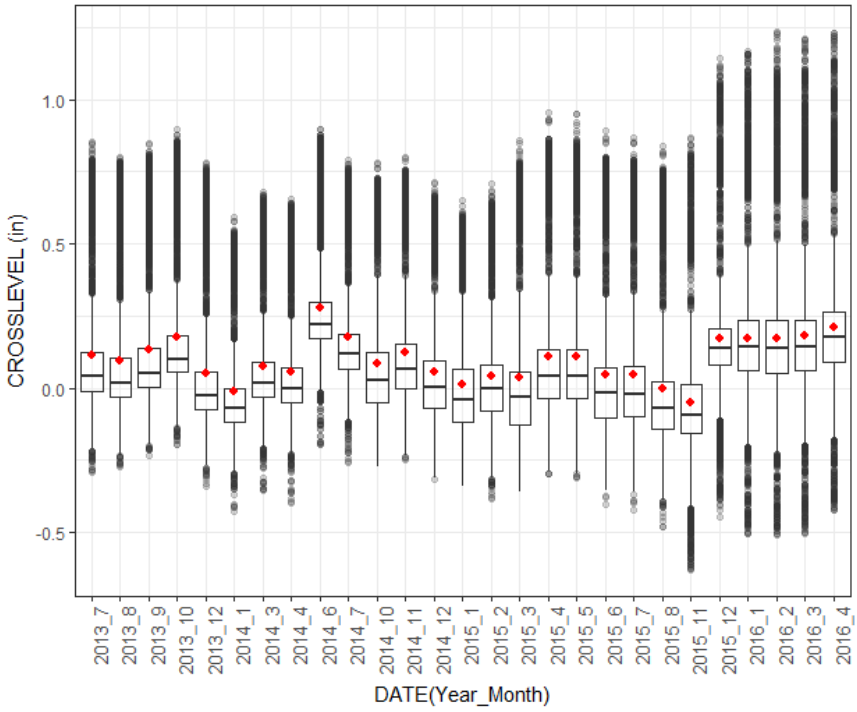


FIGURE 10 Distribution of rail cross level at measurement locations (07/2013-04/2016).

Figure 11 shows the standard deviation for crosslevel measurements at all the locations. Higher standard deviation values were observed at locations close to the end of the track section considered between 4800ft and 5200ft.

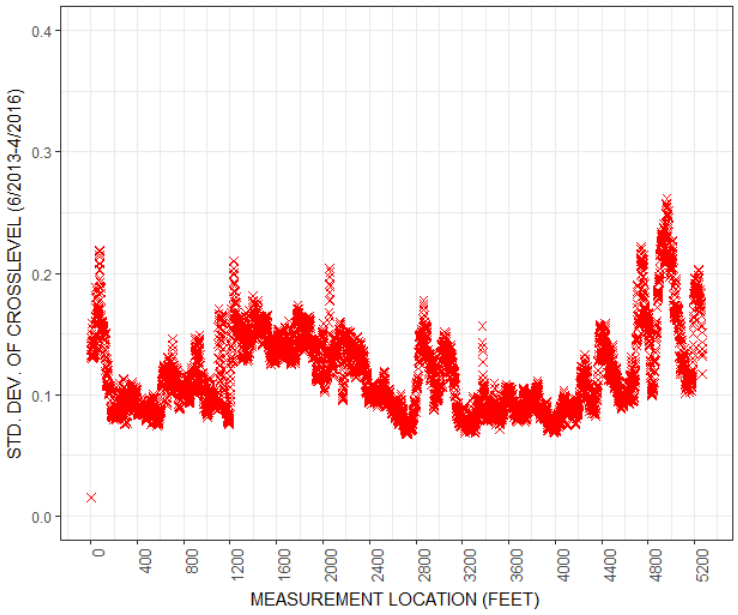


FIGURE 11 Standard deviation for crosslevel measurements at each location (07/2013-04/2016).

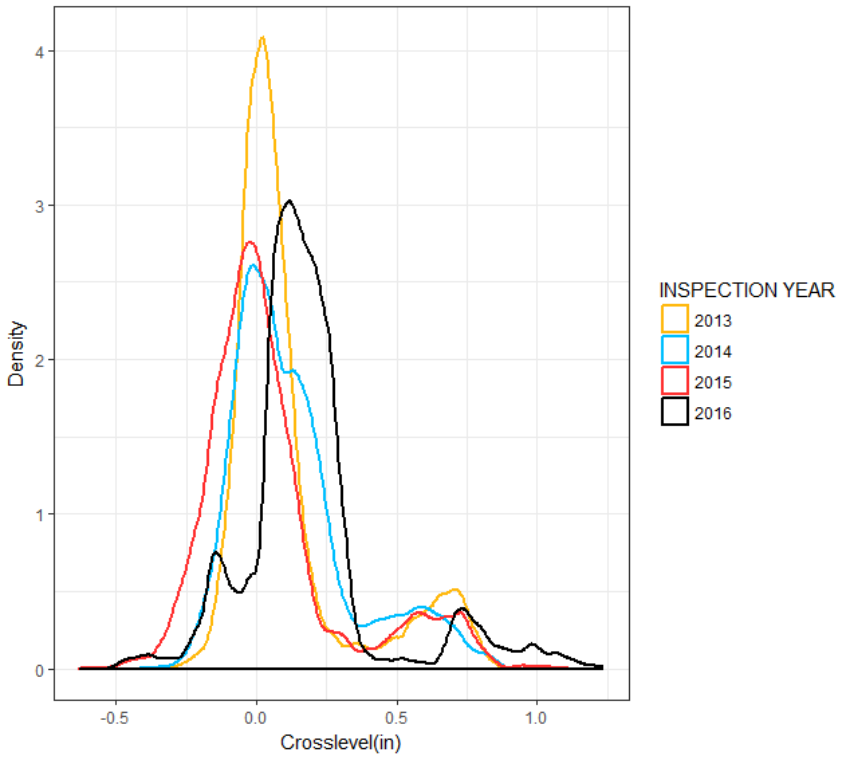
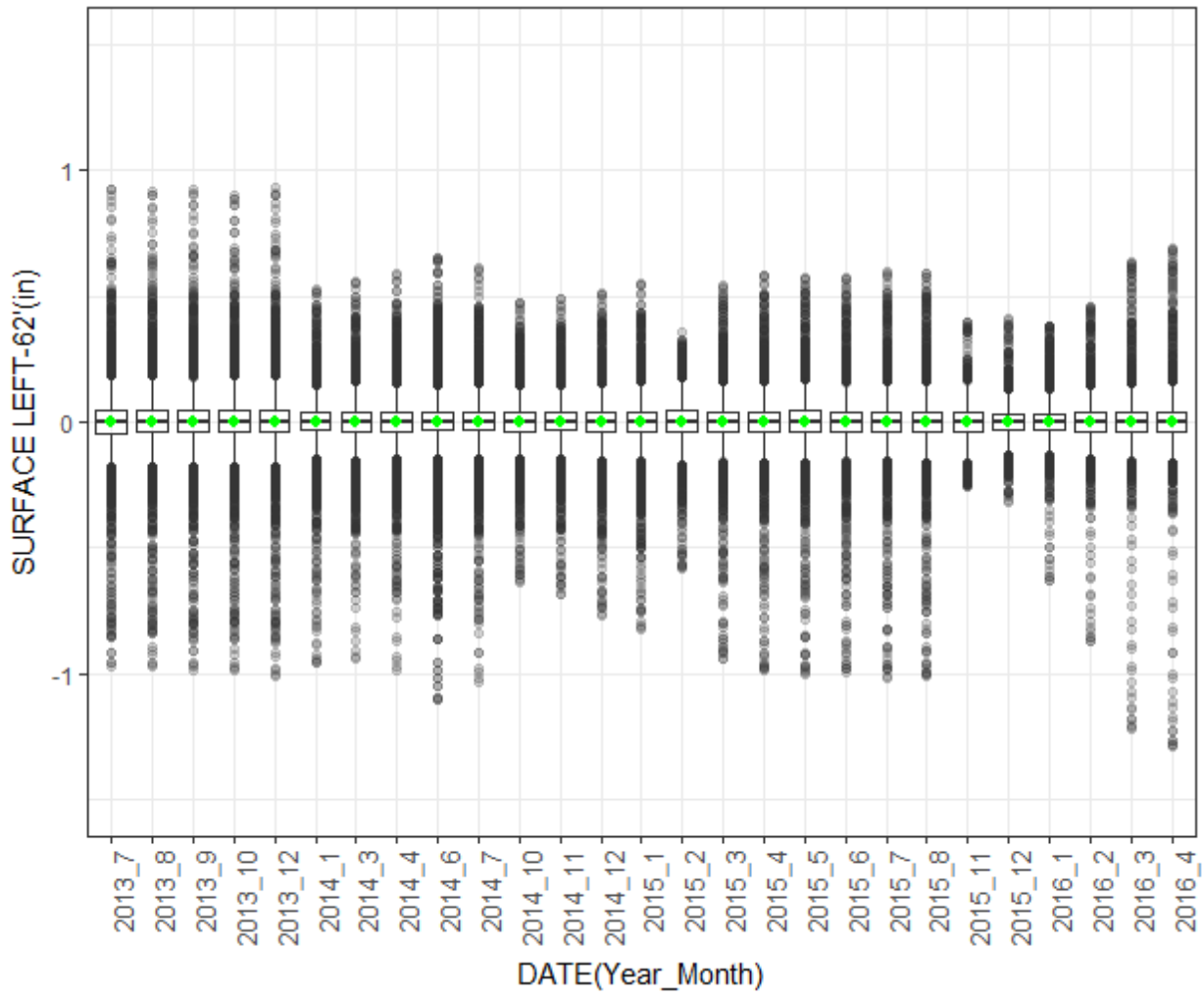


FIGURE 12 Kernel density curves for crosslevel with respect to inspection year.

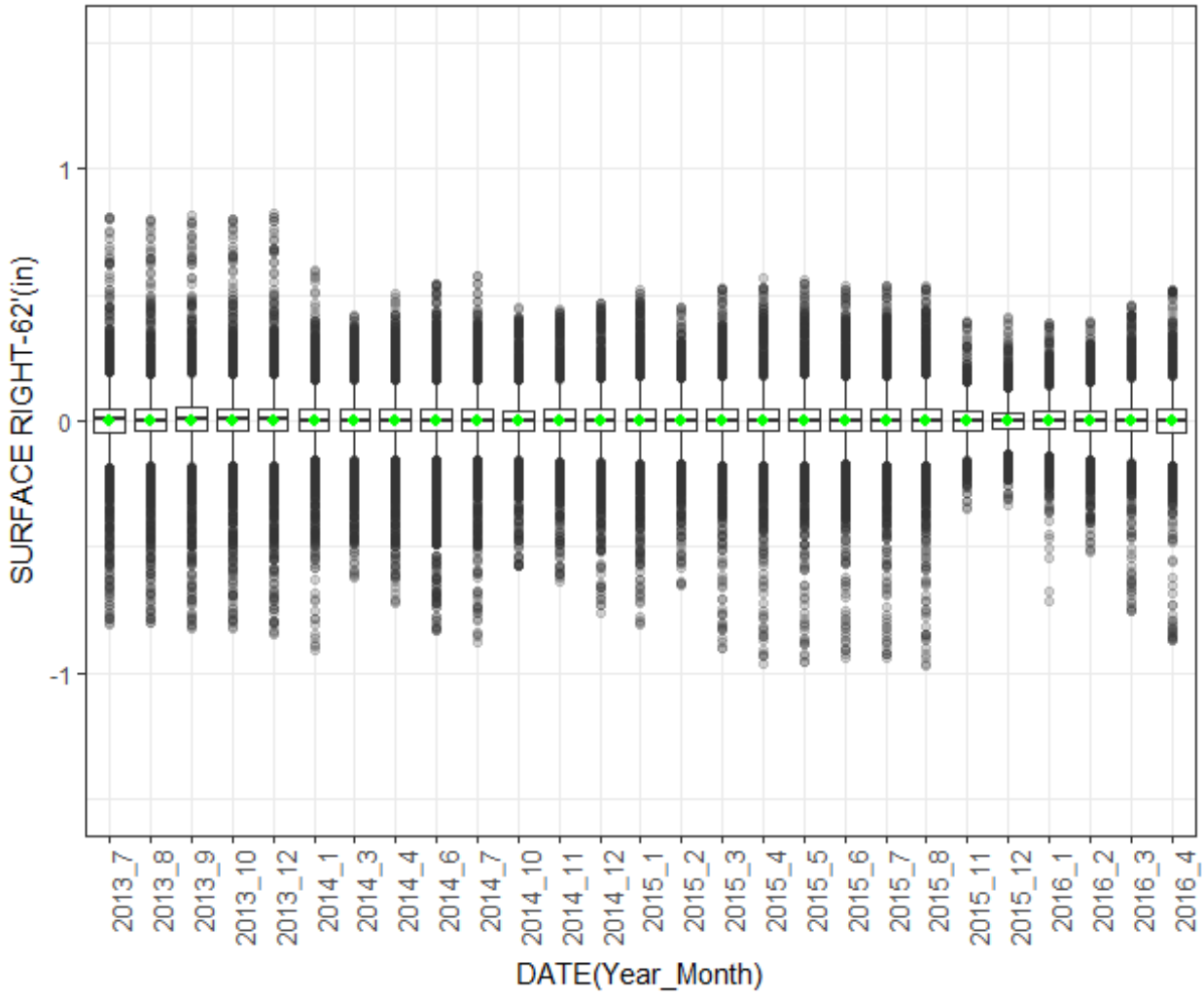
Figure 12 also shows crosslevel measurements as having a bimodal distribution. It is clear that the distribution for 2016 was different from the others considering the spread and the modal points in each.

Surface

Under ideal conditions, the surface measurements at the midordinate of a 62-foot section for this class of railroad must not exceed 1in (FRA, 2002). Figure 13 shows the surface measurements for left and right rails. Both exhibited similar behavior over time. However, there were higher observations on the left rails. There were some observations in March 2016 and April 2016 that exceeded the 1in threshold mentioned above.


SURFACE LEFT-62'(in)
1
0
-1
2013_7
2013_8
2013_9
2013_10
2013_12
2014_1
2014_3
2014_4
2014_6
2014_7
2014_10
2014_11
2014_12
2015_1
2015_2
2015_3
2015_4
2015_5
2015_6
2015_7
2015_8
2015_11
2015_12
2016_1
2016_2
2016_3
2016_4
DATE(Year_Month)

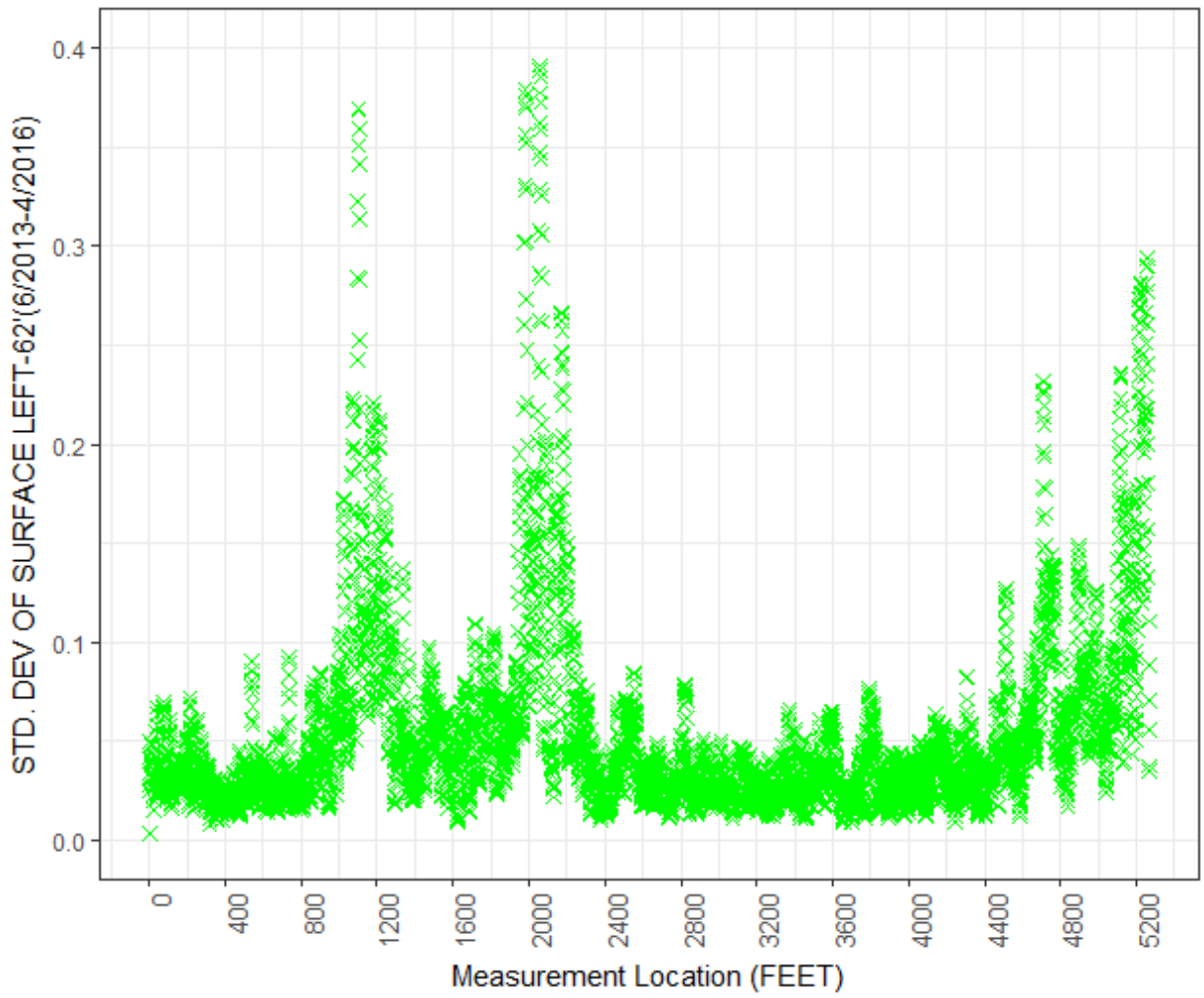
(i)



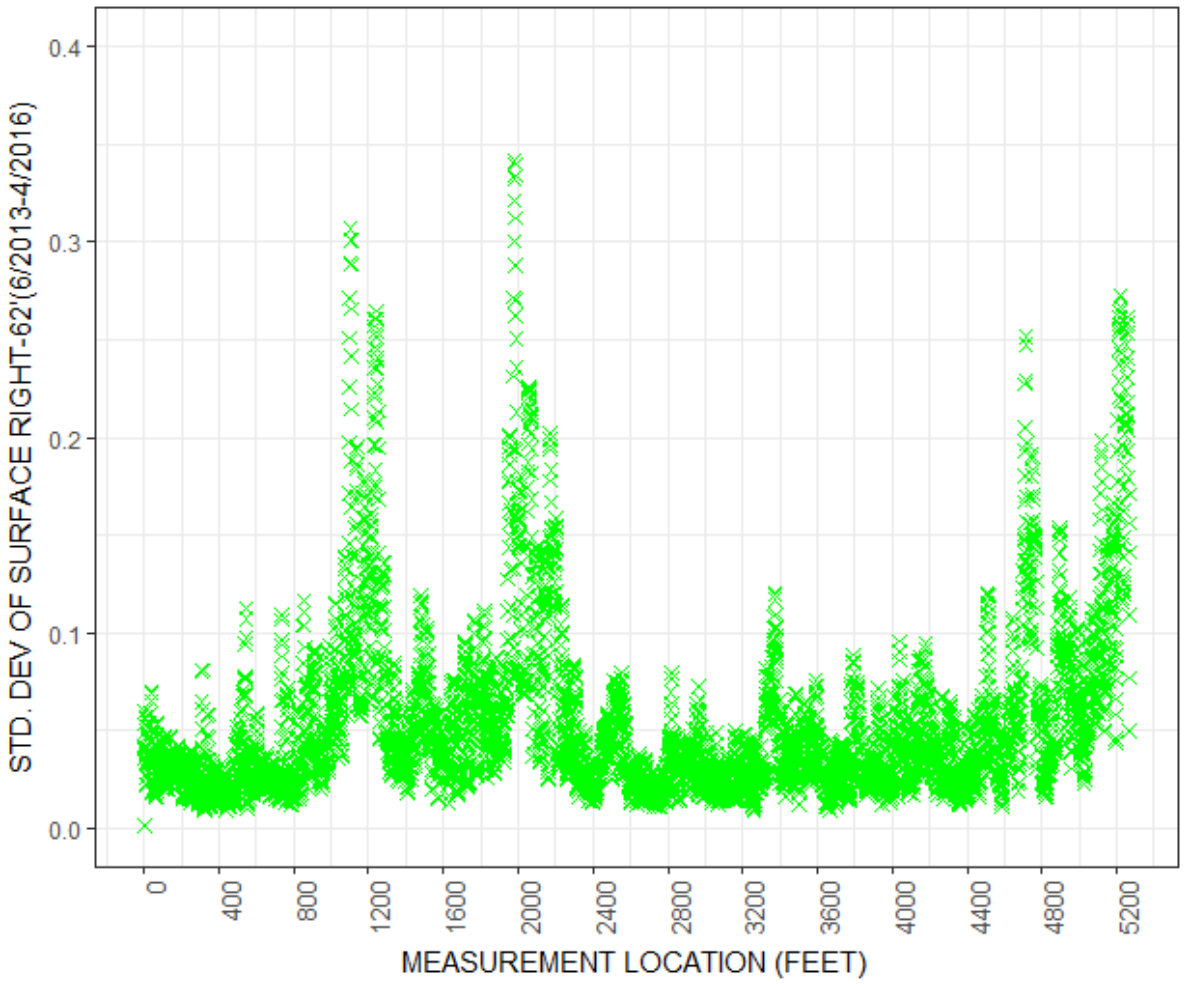
(ii)

FIGURE 13 Distribution of surface measurements: (i) Left rail (ii) Right rail (07/2013-04/2016).

Higher standard deviations for surface were observed at 1000-1200ft, 2000-2200ft, 4600-4800ft and 5000-5200ft. See Figure 14. Figure 15 shows distribution over time which remains fairly constant over the years.


STD. DEV OF SURFACE LEFT-62'(6/2013-4/2016)
0.4
0.3
0.2
0.1
0.0
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
5200
Measurement Location (FEET)

(i)



(ii)

FIGURE 14 Standard deviation for surface measurements (i) Left rail (ii) Right rail (07/2013-04/2016).

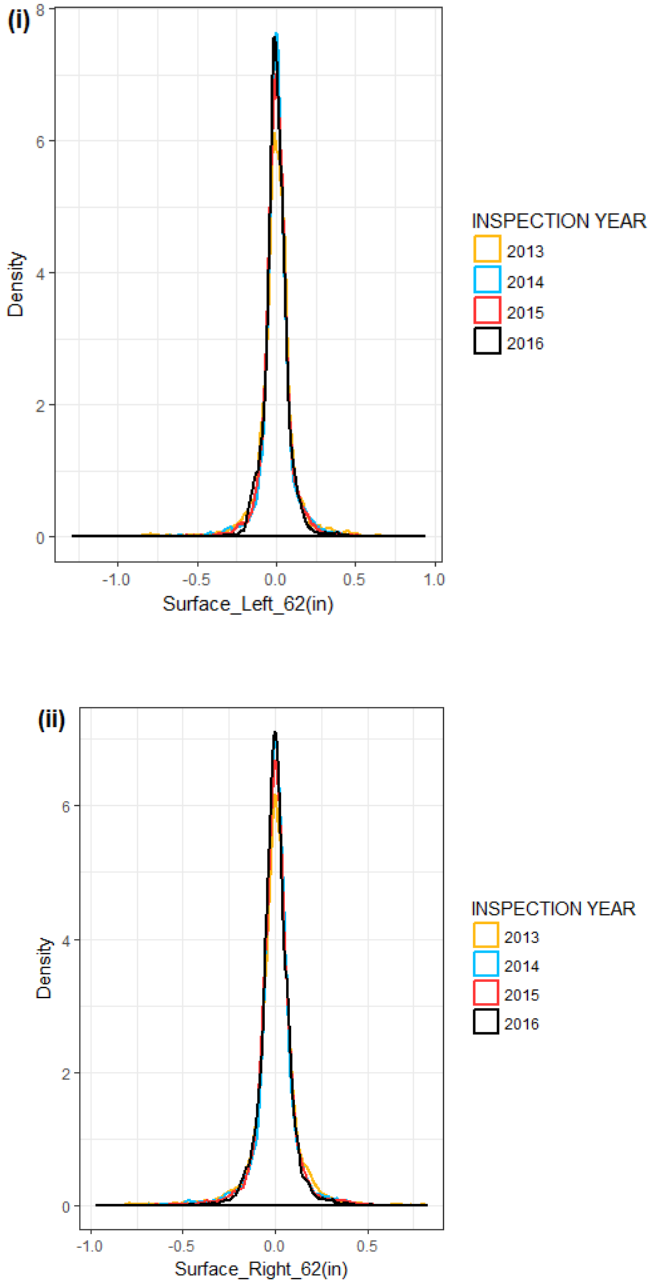
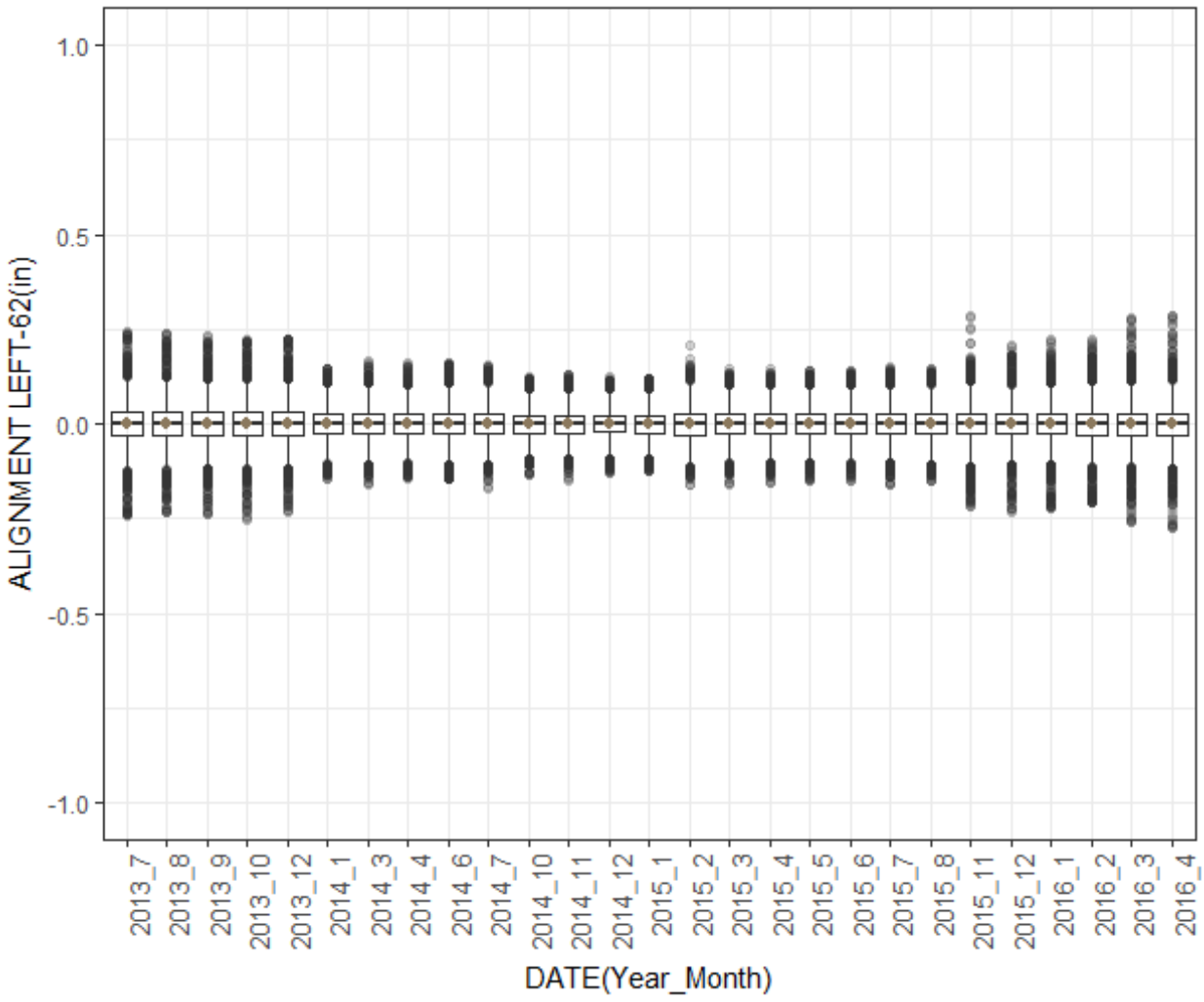


FIGURE 15 Surface measurements distribution (2013-2016).

Alignment

Alignment deviation from the midpoint of a 62-foot chord must not exceed ½" from unifomity (FRA, 2002) for this class of railroad. Figure 16 (i) and (ii) represent boxplots of alignment measures for all 5268 track locations for left and right rails within the analysis period. With the exception of April 2016, measurements for all years were below the ½" threshold.


ALIGNMENT LEFT-62(in)
1.0
0.5
0.0
-0.5
-1.0
2013_7
2013_8
2013_9
2013_10
2013_12
2014_1
2014_3
2014_4
2014_6
2014_7
2014_10
2014_11
2014_12
2015_1
2015_2
2015_3
2015_4
2015_5
2015_6
2015_7
2015_8
2015_11
2015_12
2016_1
2016_2
2016_3
2016_4
DATE(Year_Month)

(i)

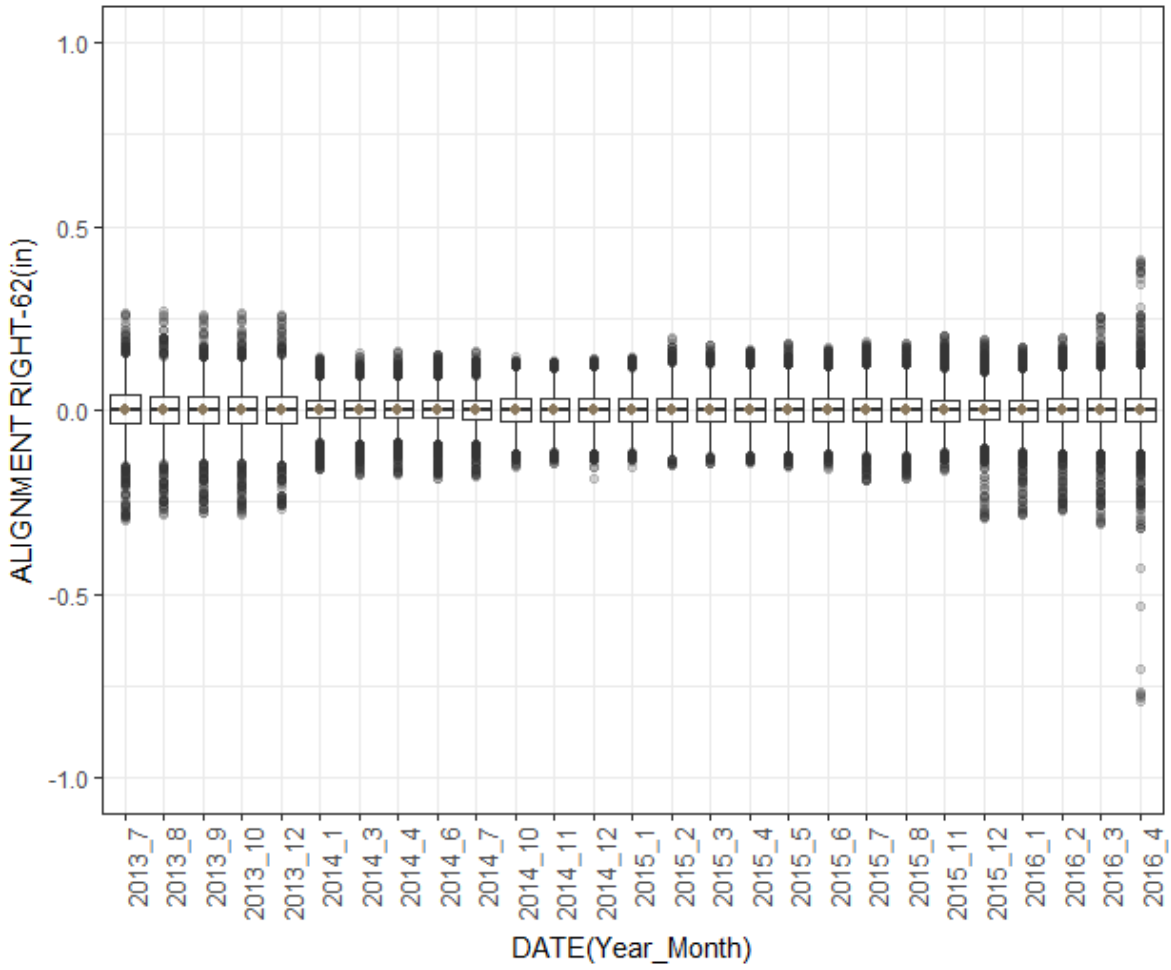
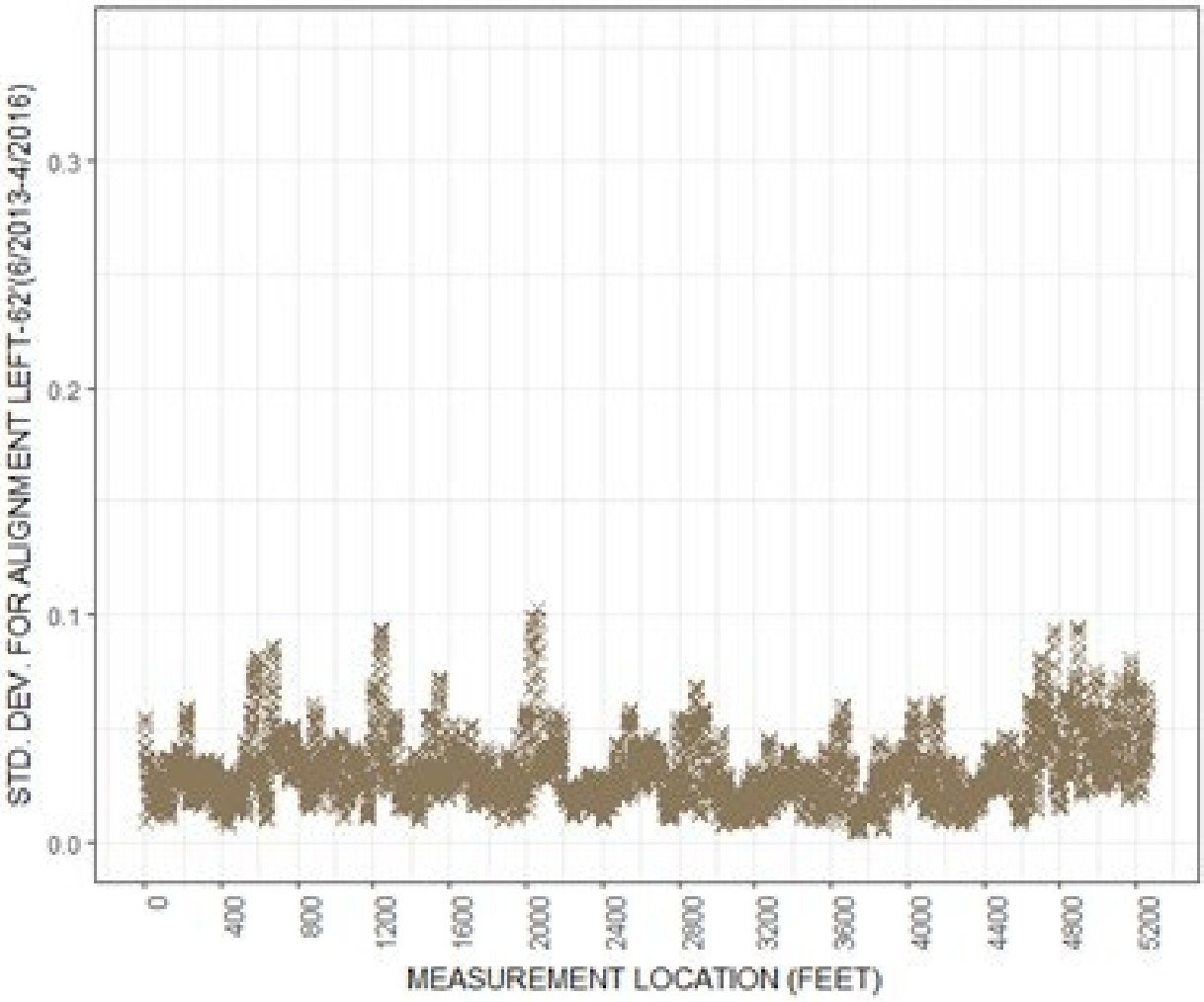
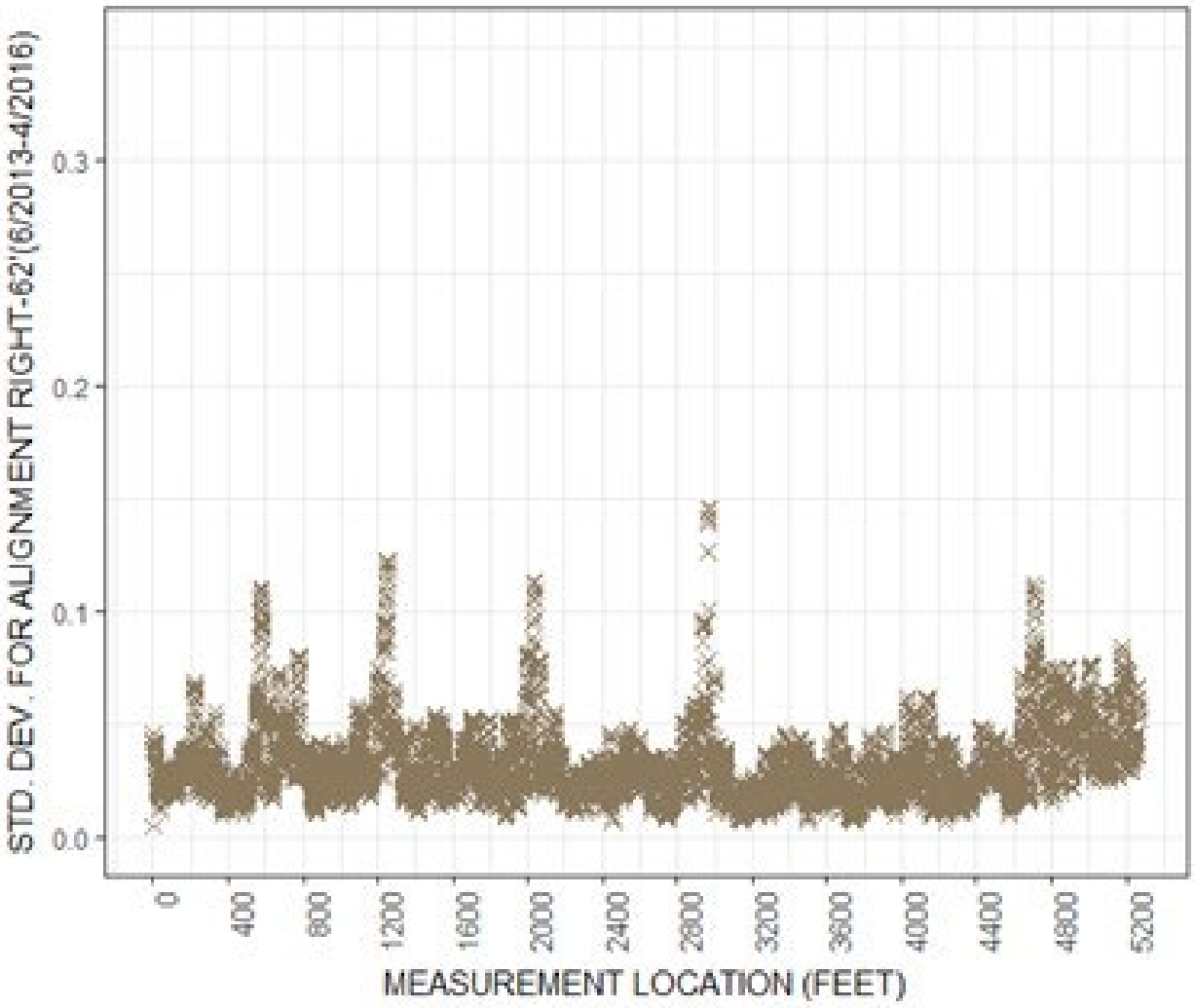


FIGURE 16 Distribution of alignment (i) Left rail (ii) Right rail (07/2013-04/2016).

The standard deviations for rail alignment at inspection locations for both rails are shown in Figure 17. Highest standard deviation for alignment were recorded on the left rail between 2800ft and 3000ft. Figure 18 is the kernel density plot for the distribution of alignment measurements.

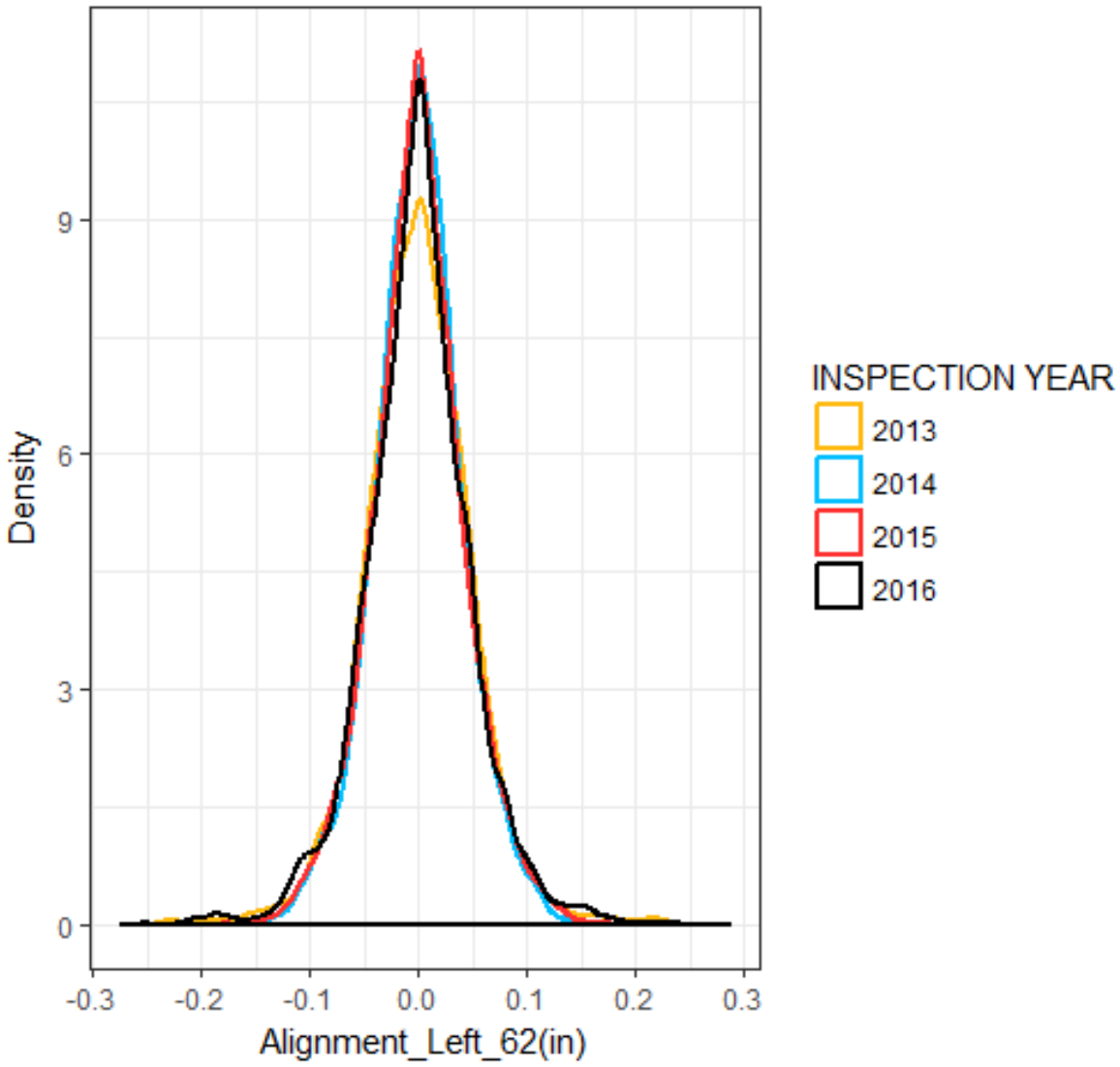

STD. DEV. FOR ALIGNMENT LEFT-62'(6/2013-4/2016)
0.3
0.2
0.1
0.0
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
5200
MEASUREMENT LOCATION (FEET)

(i)

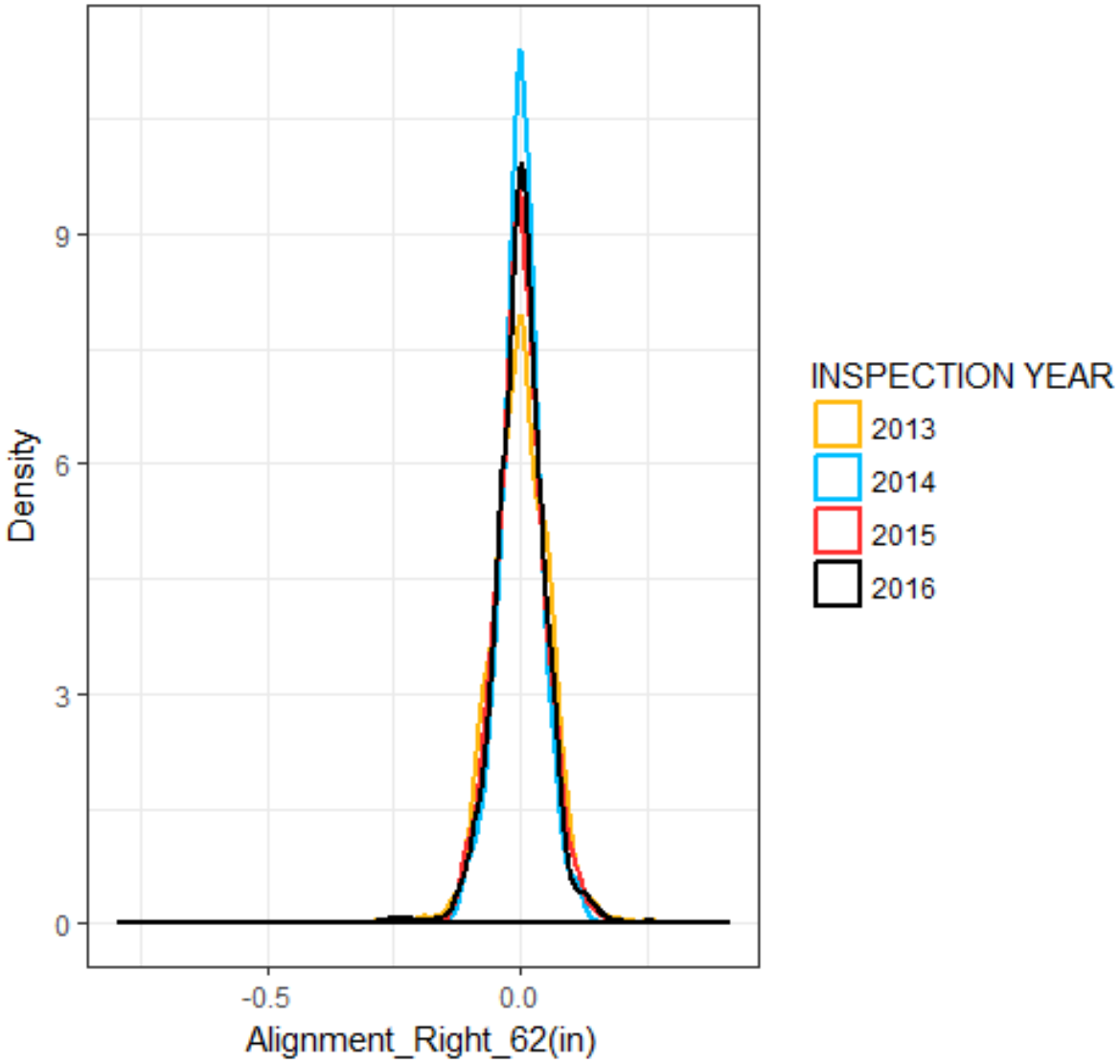


(ii)

FIGURE 17 Standard deviation for alignment (i) Left rail (ii) Right rail (07/2013-04/2016).


Density
9
6
3
0
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Alignment_Left_62(in)
INSPECTION YEAR
2013
2014
2015
2016

(i)



(ii)

FIGURE 18 Kernel density plots for alignment (i) Left and (ii) Right.

Warp

Warp is a critical safety parameter in railroads. Excessive warp can lead to wheel derailments (FRA, 2002). Difference in crosslevels between any two points must not exceed 1.5in. Figure 19 is a boxplot for all the warp measurements for the locations using a 62ft chord for all inspection dates considered in this study. Again, there was a gradually increasing trend for maximum warp values between December 2015 and April 2016. The mean warp was approximately zero throughout the analysis period.

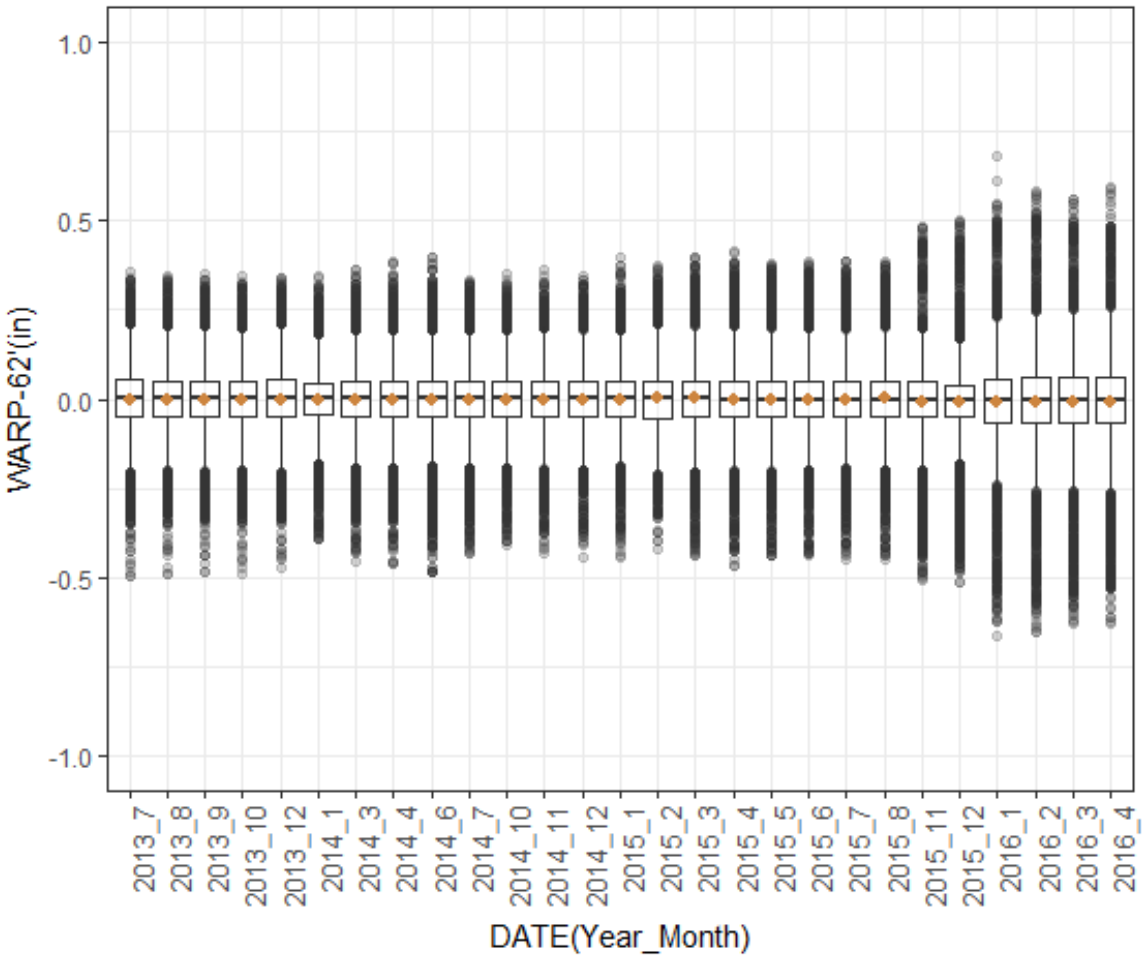


FIGURE 19 Distribution of warp measurements (07/2013-04/2016).

Again, higher standard deviations were recorded at the tail-end of the section in Figure 20. The distribution of warp measurements for 2016 was more spread out compared to the other periods as seen in Figure 21.

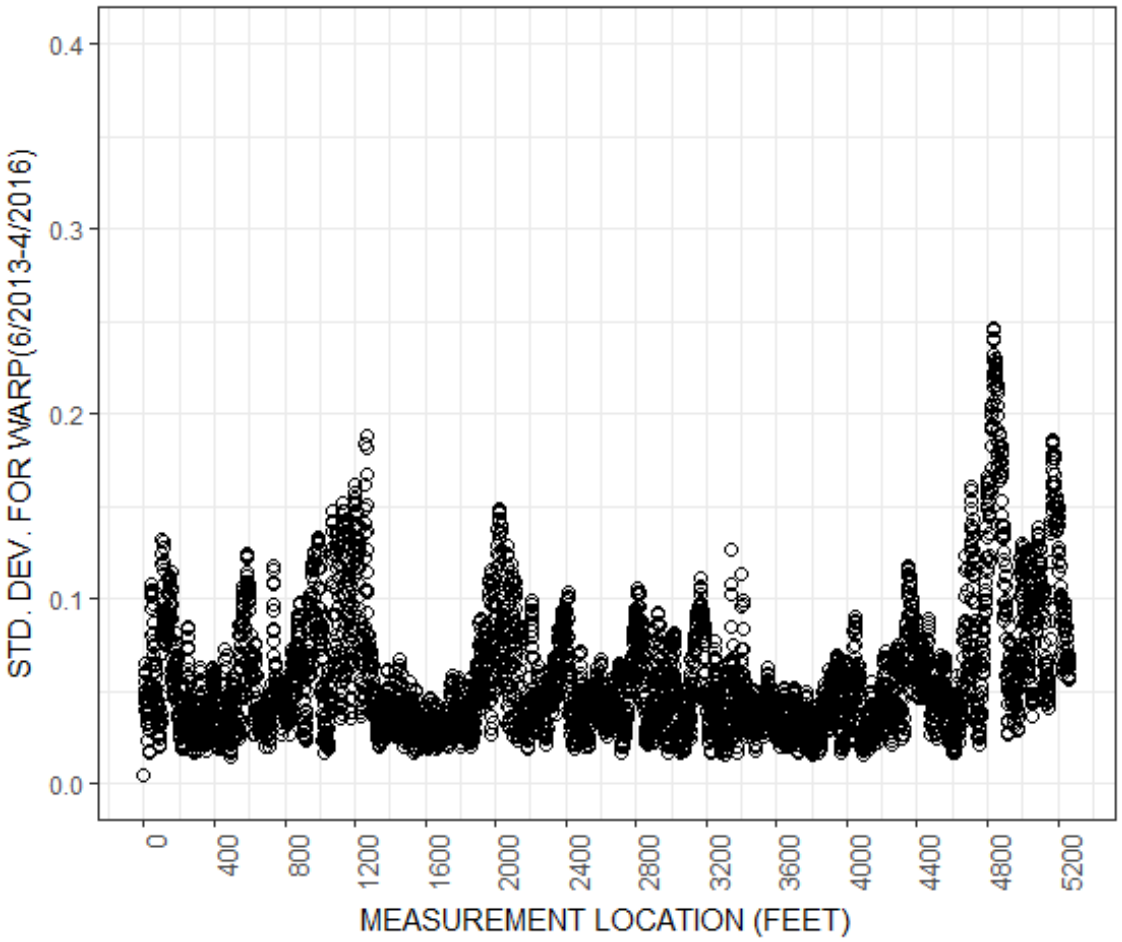


FIGURE 20 Standard deviation for warp (07/2013-04/2016).

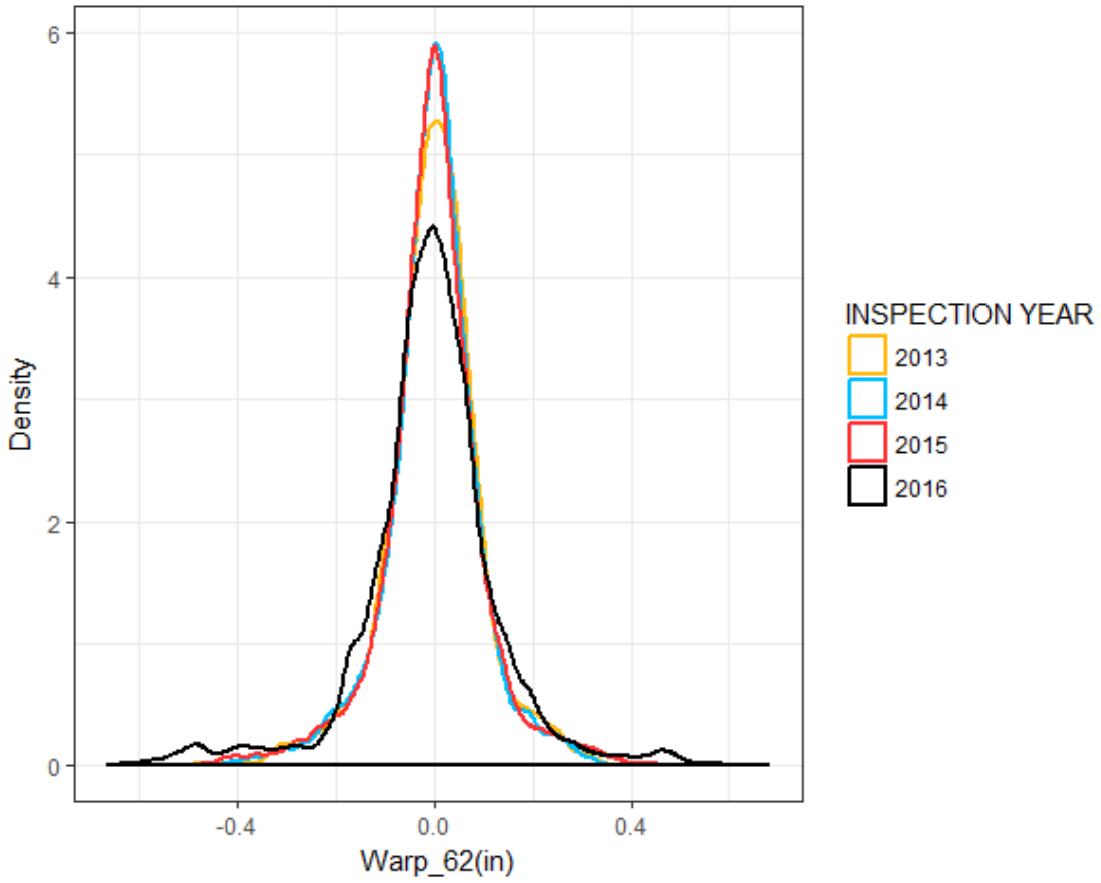


FIGURE 21 Kernel density curves for Warp with respect to year of inspection.

MULTI-WAY DATA ANALYSIS

This chapter provides details on how multiway data analysis was carried out on the rail track geometry data set.

Centering & Scaling of Data

Before tensor decomposition was carried out, track geometry data was centered and scaled. Centering and scaling eliminates unwanted differences in level and scale (Kiers, 2000). According to researchers (Bro and Smilde, 2003), centering may lead to a removal of offsets in the data and increased model fit. Centering is carried out along one mode by averaging the data along the specified mode and subtracting it from each entry along the mode. Equation 6 is the expression for centering along mode I of a tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{I \times J \times K}$. The idea of centering is also illustrated in Figure 22 below.

$$x_{ijk}^{\cdot} = x_{ijk} - \frac{\sum_{i=1}^{I} x_{ijk}}{I} \tag{6}$$

Where $x_{ijk}^{\cdot}$ is the centered data entry.

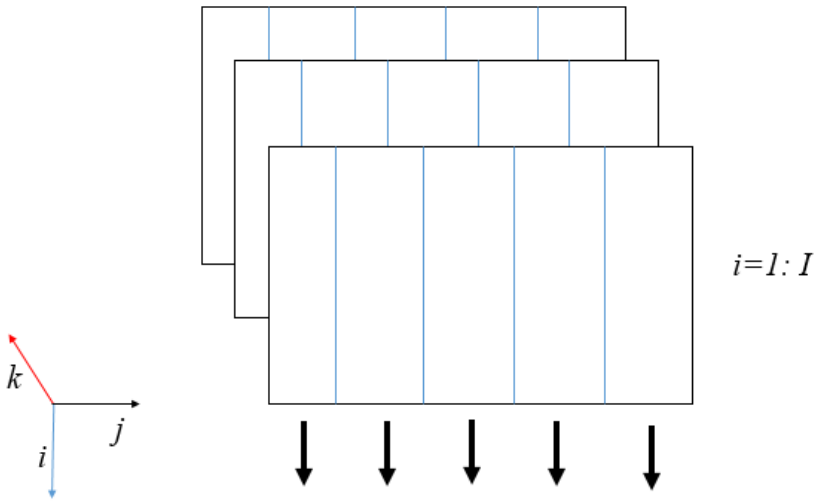


FIGURE 22 Centering along mode I.

Scaling, carried out within a mode, is performed to adjust scale differences among various parameters in the data set. Typically, scaling to unit standard deviation within the second mode is carried out leading to variables having the same variance which results in each variable having the same opportunity to influence the model (Bro and Smilde, 2003). Mathematically, scaling within the second mode of tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{I \times J \times K}$ is shown in equation 7.

$$x_{ijk}^{\cdot\cdot} = \frac{x_{ijk}}{\sqrt{\frac{\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ijk}^2}{IK}}} \tag{7}$$

Where $x_{ijk}^{\cdot\cdot}$ represents the scaled data.

PARAFAC Decomposition

PARAFAC decomposition was carried out on the track geometry data which had been centered along mode-1 and scaled across mode-2. The loading matrix for the 2nd dimension which represented the track geometry variables was constrained to be orthogonal. Orthogonality constraints were used to ensure the model captured uncorrelated underlying phenomena across the variables in the data. Table 4 below shows the various models fitted along with the core consistency diagnostic (CORCONDIAG) which is a measure of model stability. See (Andersen and Bro, 2003) and (Bro and Kiers, 2003) for detailed explanation on the core consistency diagnostic. Ideally, a value above 90% indicates a stable model describing the trilinear variation in the data, while a value close to zero suggests an invalid model since space covered by component matrices are not describing trilinear variation (Morup, 2011).

TABLE 4 PARAFAC Models

No. of components	R-squared	CORCONDIAG (%)
2	0.34	97.34
3	0.45	92.46
4	0.55	75.02
5	0.63	4.745

After 18 iterations, the 3-component model was chosen for further analysis because it explained almost half of the systematic variation in the data (45%) with a high CORCONDIAG value. Figure 23 shows 3 loading plots for different combinations of the loading factors from mode-2 loading matrix. From Figure 23 (i), it is clear that surface measurements on both tracks and gage width dominated components 1 and 2 respectively. Additionally, the consistent proximity of both surface measurements in all three plots suggests a high correlation between the two as expected. See Figure 24 which confirms the correlations. This information becomes useful when performing dimension reduction since one of the two surface measurements can be removed when modeling without severely influencing the model. Crosslevel is revealed as the dominant variable captured in component 3.

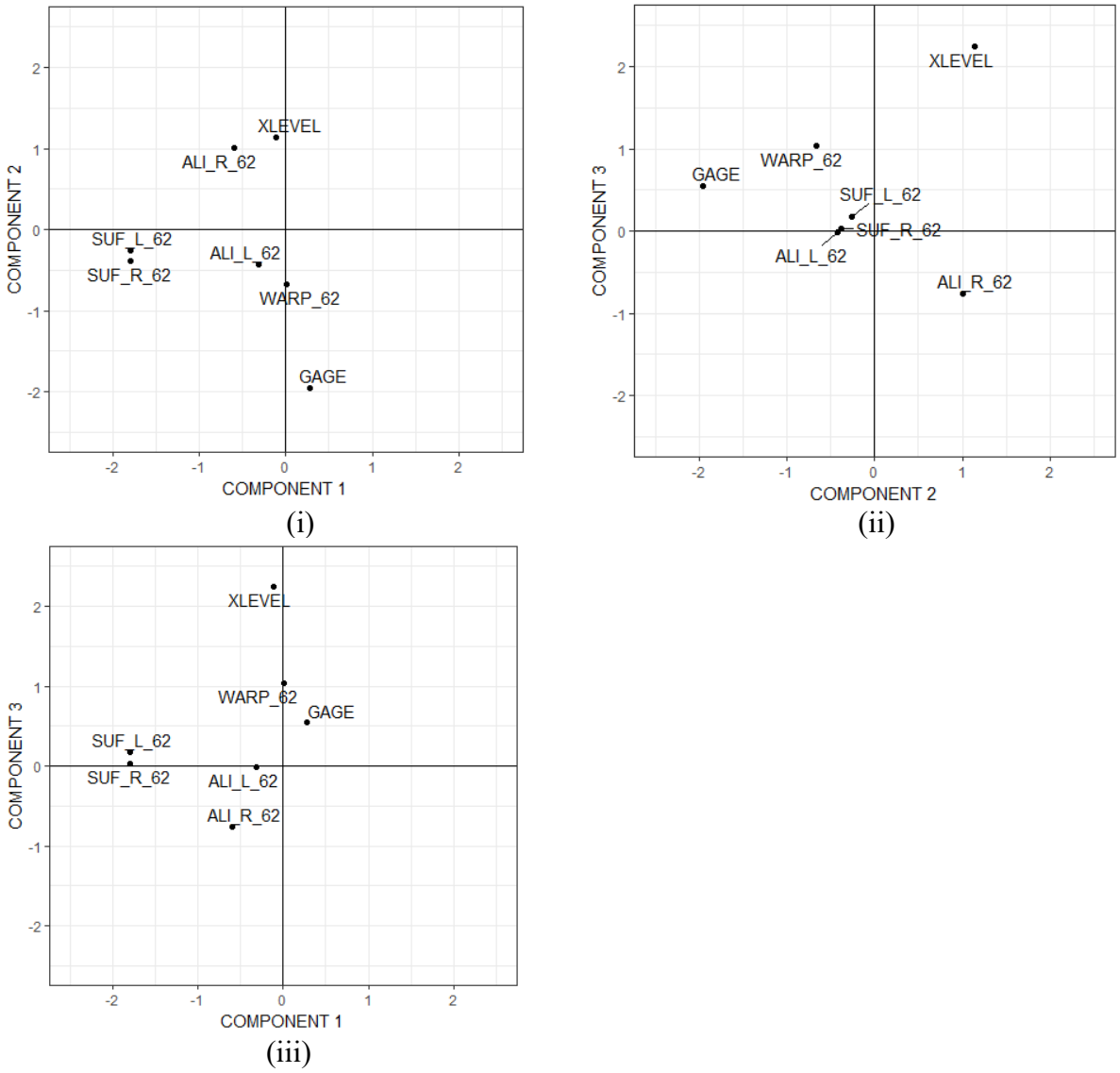


FIGURE 23 Loading plots for track geometry parameters in 3-component PARAFAC model.

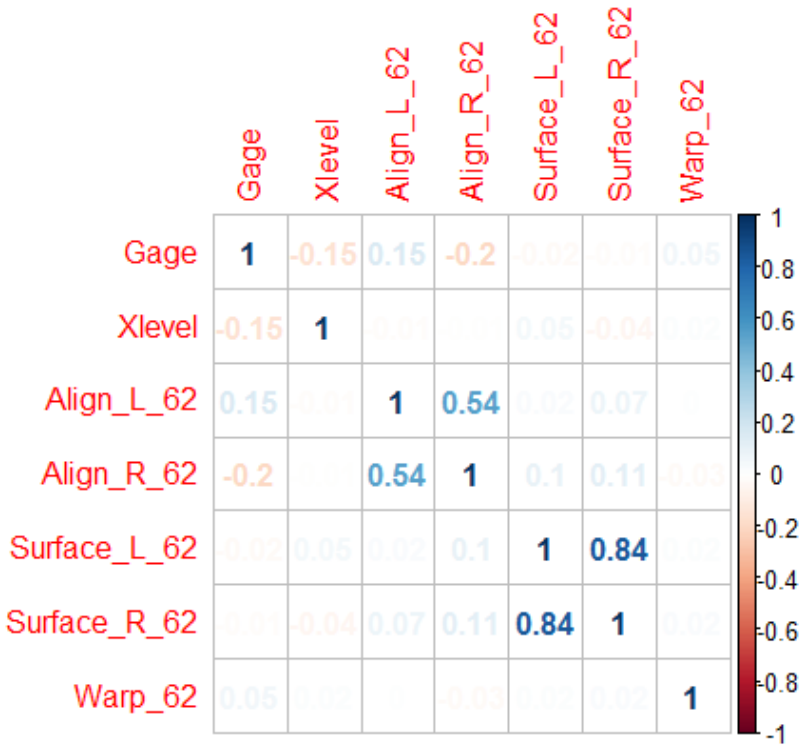


FIGURE 24 Correlation plot for track geometry parameters.

For the analysis of time variation in the data, the loading plots for inspection dates are shown in Figure 25. In Figure 25 (i), (ii) and (iii), two distinct groups of inspection dates are revealed. The first group includes the following dates: November 2015, December 2015, January 2016, February 2016, March 2016 and April 2016. The remaining inspection dates were captured in the second group. These two groups were revealed with the kernel density plots for gage (Figure 9) and crosslevel (Figure 12) during data exploration where a departure from the general trend is observed for 2015 and 2016.

Model Validation

To ensure whether the right number of components were extracted, split-half analysis is used to validate the model. This involves splitting data along into two along one mode and fitting the model. If the model has the right number of components describing the underlying behavior of the system, decomposition of the 2 halves will lead to the same results.

For this work, locations were split into two halves and the 3-component model PARAFAC model with orthogonality constraints in mode-2 is used to decompose each half of the data. Table 5 below shows the model performance for each half of the data.

TABLE 5 Results for Split-half Analysis

Data	R-squared	CORCONDIAG (%)
Half-1	0.48	97.81
Half- 2	0.46	90.96

With the split-half analysis yielding similar results to the decomposition of the entire data set, the 3-component model is suitable to explain trilinear variation in the data.

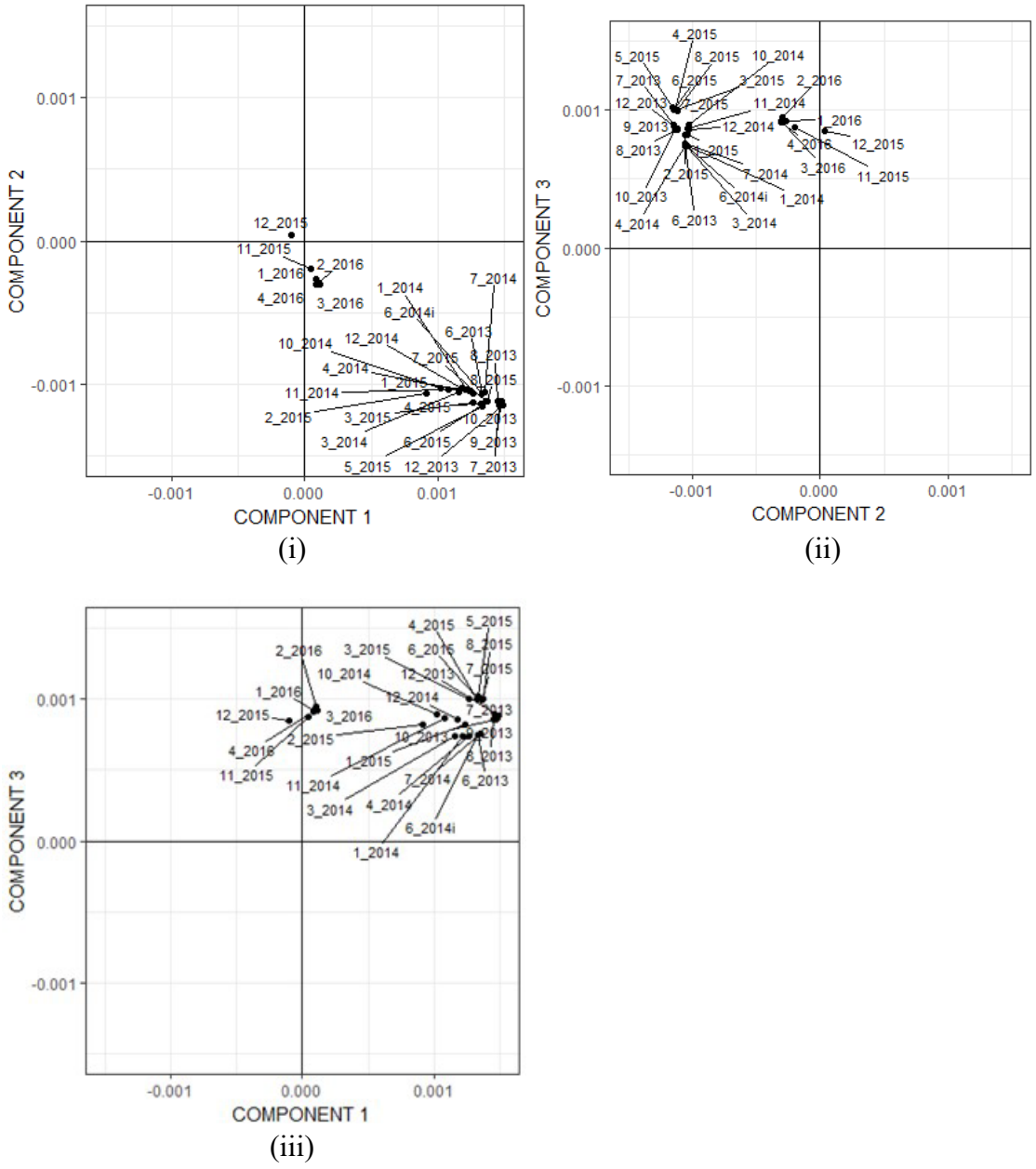


FIGURE 25 Loading plots for inspection dates in 3-component PARAFAC Model.

Comparison with a Two-Dimensional Data Analysis Approach

The 3-way model for the track geometry data was compared with a two-way model generated by the principal component analysis (PCA) to identify the benefits of a multiway analysis approach. To perform PCA, the multiway data shown in Figure 1 (ii) was flattened into a two-dimensional data set. This was achieved by averaging across the time dimension as shown in Figure 26. It must be noted that by averaging across time, information on the temporal variation of the data is lost.

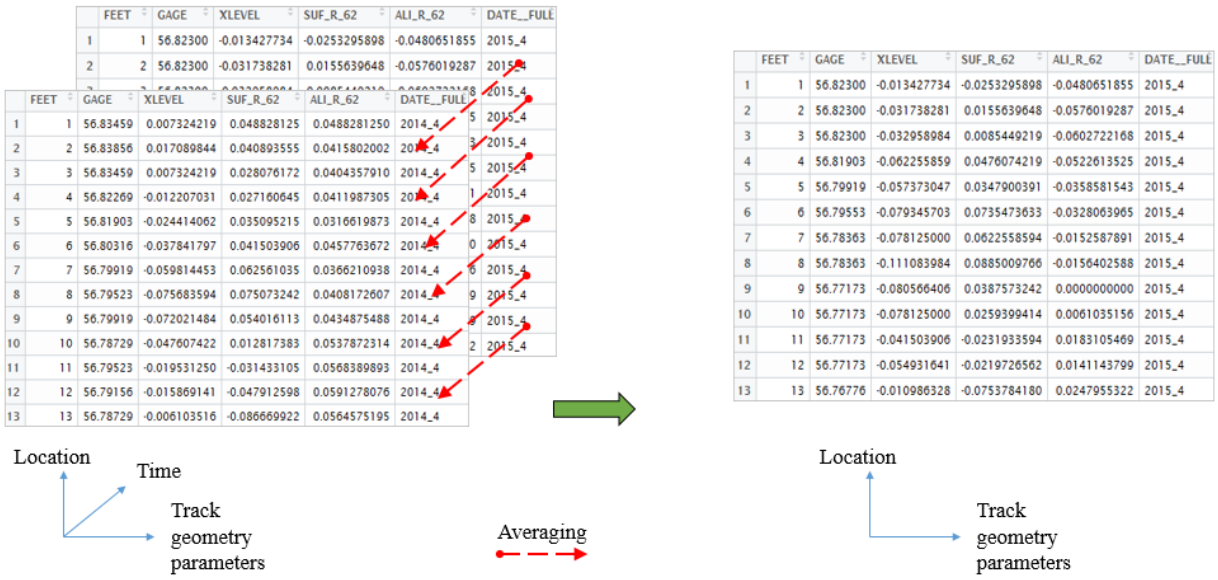


FIGURE 26 Averaging 3-dimensional data across time to obtain 2-dimensional data.

Using 50% of the locations in the data set, PCA was performed. The variance captured by each principal component is shown in Figure 27. The first and second components collectively captured 50.2% of the total variation in the data.

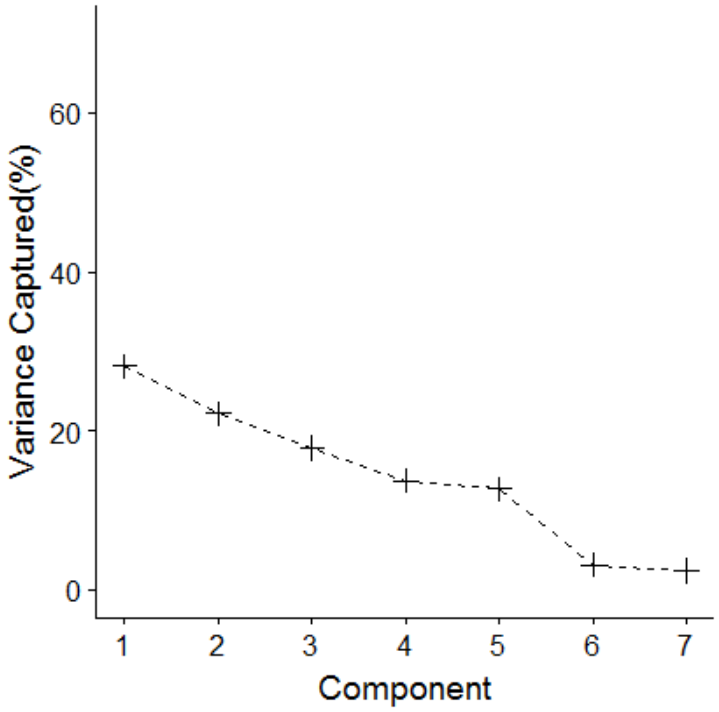


FIGURE 27 Plot of variance captured by principal components.

The biplot of the principal components is shown in Figure 28. With the exception of surface levels on the right and left tracks, the distribution of the other variables were similar to the plot of components 1 and 2 in the PARAFAC analysis in Figure 23 (i). In Figure 28 , surface level for the right and left tracks appear to be almost orthogomal suggesting that these two parameters have very low to no correlation. Considering that the track section analyzed was tangent, this observation is not valid. A look at the correlation plot (Figure 24) and loading plot (Figure 23 (i)) for the PARAFAC model clearly shows that those two parameters were in fact, the most linearly correlated among the variables.

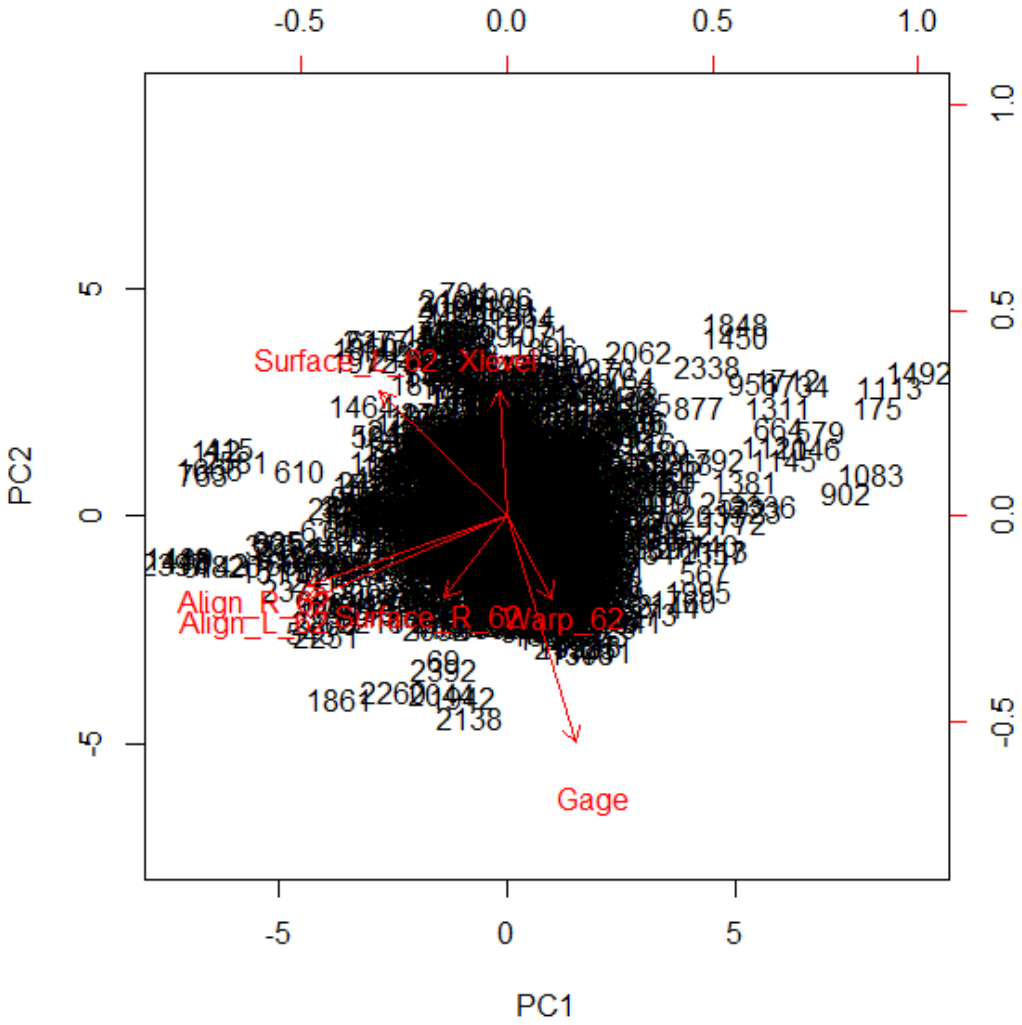


FIGURE 28 Biplot of principal components.

A look at the correlation values for surface levels on right and left rail tracks for each year in Figure 29 also confirms this. For each year, there was a consistently high correlation between the two surface measures which the PCA approach failed to capture. In effect, the three-way approach is able to capture a more accurate temporal signature of the data set compared with the 2-way approach.

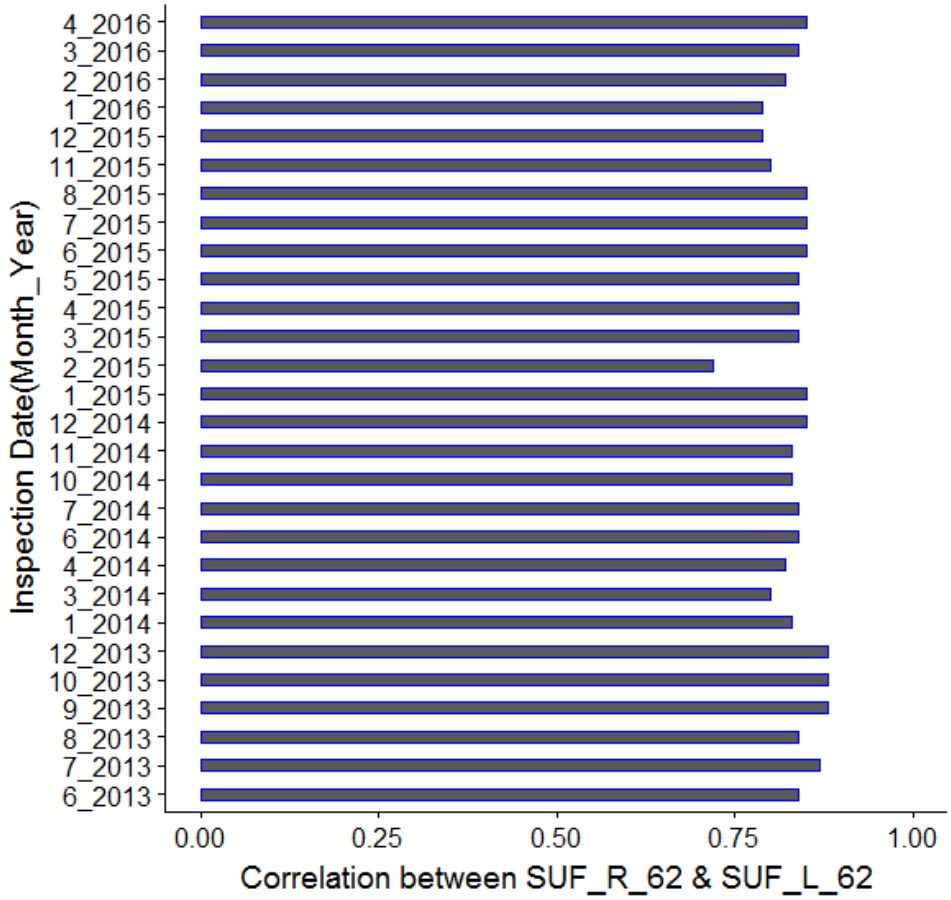


FIGURE 29 Correlation between right and left surface levels for inspection dates.

CONCLUSION

This report shows the potential benefits of using a multiway data modeling approach to analyze railroad infrastructure data collected over time. Highlights include:

- Track geometry parameters for railroad infrastructure can be considered as a three dimensional data set comprising track geometry variables measured at different locations along the track at different inspection dates
- Seven track geometry parameters considered for this study. They included: Gage, Crosslevel, Right track surface (62-foot chord), Left track surface (62-foot chord), Right track alignment (62-foot chord), Left track alignment (62-foot chord) and Warp (62-foot chord)
- The concept of multi-dimensional data analysis is suited for the data set since measurements of track geometry parameters are correlated with respect to time
- PARAFAC decomposition which is a simpler model to fit compared to the Tucker Decomposition was used to analyze the data set
- The multiway decomposition approach revealed Surface and Gage measurements as being the most dominant variables responsible for almost half of the variation in the data set
- Two distinct groups for inspection dates were also revealed by multiway analysis

- Right and left surface measurements were shown to be the most highly correlated pair implying that only one of these can be used in further modeling of the data
- PCA performed after flattening the data failed to show the high correlation between right and left surface measurements. This may have been due to the loss of temporal variation over time as a result of the averaging process to transform the data into a matrix (two-dimensional data)

Future of Multiway Data Analysis in Railroad Infrastructure

Multiway data analysis has the potential to improve railroad infrastructure management. The following are considerations moving ahead:

- Introduce other track geometry parameters to improve understanding of deterioration process and how variables are interrelated with each other
- Incorporate nonlinearity into multiway approaches to ensure nonlinear behavior of parameters are captured by multiway models
- Using multiway models as a basis for predicting future conditions of railroad track.

REFERENCES

1. Acar, E., and Yener, B. (2009). Unsupervised Multiway Data Analysis: A Literature Review. IEEE Transactions on Knowledge and Data Engineering. Vol. 21, No. 1
2. Andersen, C.M. and Bro, R. (2003). Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. J. Chemometrics 2003; 17:200-215
3. ASCE (2017). 2017 Infrastructure Report Card- Rail. https://www.infrastructurereportcard.org/wp-content/uploads/2017/01/Rail-Final.pdf Accessed: 7/20/2017
4. Bader B.W., Berry M.W., Browne M. (2008) Discussion Tracking in Enron Email Using PARAFAC. In: Berry M.W., Castellanos M. (eds) Survey of Text Mining II. Springer, London
5. Bro, R. (1997). PARAFAC. Tutorial and Applications. Chemometrics and intelligent laboratory systems 38 (1997) 149- 171
6. Bro, R. and Smilde, A.K. (2003). Centering and scaling in component analysis. J. Chemometrics 2003; 17: 16-33
7. Bro, R. and Kiers, H.A.L. (2003). A new efficient method for determining the number of components in PARAFAC models. J. Chemometrics; 17: 274-286
8. Chaolong, J., Weixiang, X., Futian, W., Hanning, W. (2002). Track Irregularity Time Series Analysis and Trend Forecasting. Discrete Dynamics in Nature and Society. Volume 2012. Article ID 387857, doi: 10.1155/2012/387857
9. FRA (2002). Track Safety Standards Compliance Manual. Federal Railroad Administration. USDOT Office of Safety Assurance and Compliance.
10. FRA (2013). Memorandum. Technical Bulletin T-13-01, Guidance regarding the application of vehicle/track interaction safety standards; high speed and high-cant deficiency operations, Final Rule, Track Classes 1-5. Federal Railroad Administration, USDOT.
11. Kiers, H.A.L. (2000). Towards a standardized notation and terminology in multiway analysis. J. Chemometrics 2000; 14:105-122
12. Kolda, T.G., Bader, B.W., Kenny J.P. (2005). Higher-Order Web Link Analysis Using Multilinear Algebra. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)
13. Kroonenberg, P.M. (2008). Applied Multiway Data Analysis. Wiley series in probability and statistics. ISBN 978-0-470-16497-6
14. Lewis, R.C. (2011). Track Geometry Recording and Usage. Notes for a lecture to Network Rail. http://www.infrastructuremonitoring.co.uk/wp-content/uploads/2011/10/Track-Recording-And-Usage.pdf Accessed: 8/9/2017
15. Meng, X., Morris, A., Martin, E. (2003). On-line monitoring of batch processes using a PARAFAC representation. J. Chemometrics 2003; 17:65-81
16. Morup, M. (2011). Applications of tensor (multiway array) factorizations and decompositions in data. WIREs Data Mining and Knowledge Discovery. Volume 1, Issue 1 p24-40.
17. Nielsen, J., Berggren, E., Lölgen, T., Müller, R., Stallaert, B., Pesqueux,L. (2013). Overview of Methods for Measurement of Track Irregularities Important for Ground-Borne Vibration. RIVAS SCP0-GA-2010-265754
18. P. Weston, C. Roberts, G. Yeo & E. Stewart (2015) Perspectives on railway track geometry condition monitoring from in-service railway vehicles, Vehicle System Dynamics, 53:7, 1063-1091, DOI: 10.1080/00423114.2015.1034730

19. Singh, K., Malik, A., Singh, V., Sinha, S. (2006). Multi-way data analysis of soils irrigated with wastewater- A case study. Chemometrics and Intelligent Laboratory Systems 83 (2006) 1-12
20. Stanimirova, I., Walczak, B., Massart, D., Simeonov, V., Saby, C.A., Crescenzo, E. (2004). STATIS, a three-way method for data analysis. Application to environmental data. Chemometrics and Intelligent Laboratory Systems 73 (2004) 219-233
21. Track Compliance Manual (2002). Track Safety Standards Classes 6 through 9. Chapter 6.
22. UFC (2008). Railroad Track Maintenance & Safety Standards- Unified Facilities Criteria (UFC). UFC 4-860-03
23. Vasilescu, M.A.O. and Terzopoulos, D. (2002). Multilinear Analysis of Image Ensembles: TensorFaces. Proc. of the European Conf. on Computer Vision (ECCV '02), Copenhagen, Denmark, p447–460
24. Zarembski, A.M. (2011). Some Examples of Big Data in Railroad Engineering. 2014 IEEE Conference on Big Data. https://pdfs.semanticscholar.org/c242/571ec25b4ecdad83e75a34a27aeeffce14b4.pdf Accessed: 8/9/2017

ACKNOWLEDGEMENTS

The authors wish to thank and acknowledge the US Department of Transportation, University Transportation Center Program (RailTEAM UTC) for funding support for this research.

ABOUT THE AUTHORS

Offei Adarkwa, Ph.D.

Mr. Offei Adarkwa is an Asset Management Engineer at The Kercher Group, Inc. He graduated from the University of Delaware with a Ph.D. in Civil Engineering in 2015. His Ph.D. work focused on the use of tensor decomposition as a data analysis tool for civil infrastructure systems, specifically bridges. His research interests include equity & debt investment vehicles for infrastructure assets, private participation in infrastructure development and transportation asset management. Dr. Adarkwa obtained his BS degree from Kwame Nkrumah' University of Science and Technology, Kumasi, and his MS degree from the University of Delaware.

Nii O. Attoh-Okine, Ph.D., P.E., F. ASCE, Snr Member IEEE

Dr. Nii O. Attoh-Okine is Professor of Civil and Environmental Engineering, and Electrical and Computer Engineering. He is also the Interim Academic Director of the University of Delaware Cybersecurity Initiative. In the last couple of years, he has authored two books which are defining the direction of research across disciplines: a) Resilience Engineering: Models and Analysis and b) Big Data and Differential Privacy in Railway Track Engineering. He is a founding associate editor for ASCE/ASME Journal of Risk and Uncertainty Analysis. He has served as an Associate Editor on the four ASCE Journals. Attoh-Okine is currently a member of a group of researchers from the United States and Japan working on Smart Cities and various cyber issues related to the Tokyo 2020 Olympic Games.