Transportation Research Part C 91 (2018) 230–248

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Principal components analysis and track quality index: A machine learning approach

Ahmed Lasisi, Nii Attoh-Okine*

Department of Civil and Environmental Engineering, University of Delaware, Newark, DE, USA

ARTICLE INFO

Keywords:
Track quality index
Principal components
Machine learning
Railway engineering

ABSTRACT

Track geometry data exhibits classical big data attributes: value, volume, velocity, veracity and variety. Track Quality Indices-TQI are used to obtain average-based assessment of track segments and schedule track maintenance. TQI is expressed in terms of track parameters like gage, cross-level, etc. Though each of these parameters is objectively important but understanding what they collectively convey for a given track segment often becomes challenging. Several railways including passenger and freight have developed single indices that combines different track parameters to assess overall track quality. Some of these railways have selected certain parameters whilst dropping others. Using track geometry data from a sample mile track, we demonstrate how to combine track geometry parameters into a low dimensional form (TQI) that simplifies the track properties without losing much variability in the data. This led us to principal components. To validate the use of principal components as TQI, we employed a two-phase approach. First phase was to identify a classic machine learning technique that works well with track geometry data. The second step was to train the identified machine learning technique on the sample mile-track data using combined TQIs and principal components as defect predictors. The performance of the predictors were compared using true and false positive rates. The results show that three principal components were better at predicting defects and revealing salient characteristics in track geometry data than combined TQIs even though there were some correlations that are potentially useful for track maintenance.

1. Introduction

This paper examines the potential of machine learning applications in railway track engineering. The railroad industry across the world is experiencing an increased demand in operations and services due to the world's increasing human needs and global trade advancements. The US Class I freight railroads alone is worth over $60 billion operating approximately 140,000 miles of track (FRA, 2013). This is second to none in the world. Other major railway networks include but not limited to Russia, China, Japan, etc. are also undergoing similar growth. The United Kingdom currently boasts of about 20,000 miles of track which currently ranks it in the first 20 rail networks in the world (Odlyzko, 2016). The global railway network is expected to reach an estimated 900,000 miles-route in 2020 (Statista, 2017). As promising as the industry is, a significant amount of the industry's budget is continuously committed towards repairs and maintenance. The annual capital program for US railroads alone is about $15 billion (Zarembski, 2015). A railway disaster is often very ruinous and sometimes fatal. Track geometry-caused derailments constitute a significant portion of railway accidents worldwide. In the United States alone, track-caused accidents have consistently constituted 30–40% of total

This article belongs to the Virtual Special Issue on "Big Data Railway".

* Corresponding author.
E-mail addresses: aolasisi@udel.edu (A. Lasisi), okine@udel.edu (N. Attoh-Okine).

https://doi.org/10.1016/j.trc.2018.04.001
Received 26 July 2017; Received in revised form 14 March 2018; Accepted 2 April 2018

accidents for the past decade (FRA, 2015). This is only second to human as a primary cause category.

Understanding track geometry from a design or mechanistic perspective is often not far-fetched. However, it is difficult to monitor the performance of track infrastructure and geometry throughout its life cycle with the same mechanistic models used during design without considering inevitable influences of the loading environment, climate, track-train dynamics and a host of other factors (DYK, 2014). To continuously, assess track geometry performance, it is common practice to take 'per-foot' measurements of crucial track geometry parameters for every mile of track using track geometry cars (Kelshaw, 1995). With varying geometry parameters and wavelength complexities, measurements taken on every foot of mile across 140,000 miles of track in the US results in about a billion observations for a given inspection date. If monthly routine inspections are scheduled, the data explodes to about several billion observations. Other measurement systems include high resolution track imaging, wheel impact load detection, strain gauge measurements, LIDAR and GPR-based ballast measurements, rail wear inspection system, etc. (Zarembski et al., 2013). The scale of data measurements in US railroads has grown from megabytes and gigabytes in the past three decades to terabytes and petabytes today. With this 'mess' of perpetual data collection, we are confronted with two major challenges that are characteristic of a typical 'big data' problem (Ishwarappa and Anuradha, 2015). Firstly, the problem of storage has been significantly confronted by modern cloud solutions. Secondly, data analysis and processing time remains a major challenge in order to make sensible conclusions from the entire ocean of data collected. Therefore, the fact that track geometry data exhibits classical big data attributes cannot be overemphasized.

In this paper we investigate the possibility of reducing multivariate track geometry indices into a low-dimensional form without losing much information. Similar to the Pavement Condition Index in highways wherein weights are assigned to each parameter and then summed up (Karim et al., 2016). However, our proposed approach takes cognizance of the fact the observed multidimensional data often lies in an unknown subspace of two to three dimensions (Hastie et al., 2009). Hence, detecting this subspace in track geometry data can significantly enable us to eliminate redundant information. This will make it possible to visualize multidimensional track geometry data in two or three dimensions which was hitherto impossible with the raw parameters obtained from track geometry cars. The second section of this paper focuses on introducing track geometry parameters, data collection and track quality indices. The third section considers selected machine learning methods that are used to train, test and validate the use of single and combined track quality indices including the proposed principal components. Low-dimensional representation of multivariate track geometry parameters in terms of principal components was validated and compared to existing TQIs in the penultimate section. The last section of this paper highlights key findings with concluding remarks.

1.1. Overview of track geometry derailments

A major component of railway accidents and derailments are due to track geometry irregularities. These irregularities include wide gage, excessive warp/twist, horizontal and vertical rail deformities which could also culminate in broken rails and derailments. The Federal Railroad Administration (FRA) reported over 450 track geometry-caused derailments in 2016 according to major cause categories. In the preceding year, the worth of track geometry derailments constituted about 15% of all major accidents (FRA, 2015). While derailments are often misunderstood to imply all train accidents, there are in fact other train accidents but statistics show that a significant proportion of accidents are derailments (Scientific American, 2016). As a result, major railways incur huge expenditures on track maintenance activities alone. These activities include inspection programs, rail grinding, tamping, etc. so as to reduce the risk of track geometry failure. In 2008 alone, about $8 billion was spent by US Class I railroads on track maintenance activities (Peng, 2011). A substantial portion of this cost go into geometry inspection. Historically, track geometry inspections were carried out manually. In recent times, the use of automated inspection cars to collect and assess track geometry cars has significantly expedited the inspection process spatially and temporally. The increase in the deployment of track geometry cars and several cutting-edge technologies has aided the performance of risk-based scheduling of track maintenance activities with the philosophy to simultaneously minimize cost, optimize operations and reduce risk (Jamshidi et al., 2017).

1.2. Background

In the area of machine learning and track geometry data analytics, some researchers have focused on track classification, track structure, components and defect detection (Molodova and Li, 2014). Significant contributions have also been made in the area of predictive maintenance and condition monitoring. Yella et al. (2009) employed a classifier fusion to categorize sleepers in railway tracks using machine vision. Using support vector machine and classifier fusion, the researchers arrived at a prediction accuracy of 92%. Fernando et al. (2011) worked on a machine vision inspection of railroad track sorting different track components from one another. They outlined an algorithm using edge detection and texture information to inspect fasteners, cut spikes and switch components. In the area of geometry forecasting, Peng (2011) proposed a short-range prediction method to forecast the amplitude of future measurements by track geometry cars. With Bayesian statistics, a work which focused on a classification-based learning using a tree augmented naïve Bayes track quality index (TAN-TQI) established better prediction accuracy than the short-ranged method (Bai et al., 2016) used to predict track irregularity of each day in a future short period (Xu et al., 2012). Hu and Liu (2016) completed a recent work on modelling track geometry degradation using support vector machine. The trained classifier was 70% accurate when tested on a validation set. Understanding the behavior of track geometry data is an important step towards developing an effective classification or regression tool. Investing some resources in this area prior classification would greatly improve on the results of the above-cited papers. Li et al. (2008) conducted some correlation analysis between track geometry and surface defects like squats. The research found a close correlation between short-wave track irregularities and the latter. Whether track quality is an indicator of