Hybrid Reduction Techniques With Covariate Shift Optimization in High-Dimensional Track Geometry

Ibrahim Balogun
Department of Civil and Environmental Engineering,
University of Delaware,
301 Dupont Hall,
Newark, DE 19716
e-mail: iobalo@udel.edu

Nii Attoh-Okine[1]
Professor
Department of Civil and Environmental Engineering,
University of Delaware,
301 Dupont Hall,
Newark, DE 19716
e-mail: okine@udel.edu

In discussions of track geometry, track safety takes precedence over other requirements because its shortfall often leads to unrecoverable loss. Track geometry is unanimously positioned as the index for safety evaluation—corrective or predictive—to predict the rightful maintenance regime based on track conditions. A recent study has shown that track defect probability thresholds can best be explored using a hybrid index. Hence, a dimension reduction technique that combines both safety components and geometry quality is needed. It is observed that dimensional space representation of track parameters without prior covariate shift evaluation could affect the overall distribution as the underlying discrepancies could pose a problem for the accuracy of the prediction. In this study, the authors applied a covariate shift framework to track geometry parameters before applying the dimension reduction techniques. While both principal component analysis (PCA) and t-distributed stochastic neighbor embedding (TSNE) are viable techniques that express the probability distribution of parameters based on correlation in their embedded space and inclination to maximize the variance, shift distribution evaluation should be considered. In conclusion, we demonstrate that our framework can detect and evaluate a covariate shift likelihood in a high-dimensional track geometry defect problem.
[DOI: 10.1115/1.4051597]

Keywords: covariate shift, derailment, collision, PCA, TSNE, track geometry

1 Introduction

Railway infrastructure is arguably the frontline asset of a country for passenger and freight transportation. Technically, to better optimize the infrastructure's performance, maximum attention must be paid during the pipeline stages: design, operation, and maintenance [1]. Additionally, rail transportation also decongests the usual traffic flow, contributes fewer gases to the atmosphere, and can transport an appreciable number of passengers [2]. Despite regular maintenance and safety rules, the havoc wrought by the failure of track geometry parameters runs into billions of dollars a year. According to the National Transportation Safety Board (NTSB), approximately every 2 h, a person or vehicle is hit by a train in the United States (US) as a result of the derailment, train-train collisions, train-car collisions, and train-person collisions [3]. Records show that the US rail network consists of about 155,000 miles of operating routes [4], and the number of accidents for the past 5 years shown in Fig. 1 corroborates the NTSB claims [5,6]. Additionally, most railway accidents are identified with three major categories: rail equipment, highway-rail grade crossing incidents, and fatalities [7].

On average, in the European Union (EU), one person is killed, and one is seriously injured at railway level crossings each day [8]. Many railroad problems have been solved with adequate geometry parameterization; however, some models' accuracy tends to degrade as the geometry data grows [9]. A simple way to minimize these losses is by adhering to efficient maintenance planning that intuitively controls track quality and restores possible off-threshold track parameters.

In the US, the annual maintenance cost for railroads is estimated to be $1.5 billion [10]. Similarly, European countries expend about 25 billion EUR annually to maintain approximately 300,000 km of track [11]. Anecdotally, we can conclude that most developed countries invest heavily in railroad maintenance to restore smooth ridership, considering threshold limits. To proffer long-term solutions, we consider the geometry problem with the individual parameter thresholds shown in Table 1 and verify the variational distribution among the geometry parameters using the covariate shift adaptation technique.

Research on covariate shift investigation on track geometry data has not received as much attention in transportation safety literature as other aspects of railroad safety. For example, various machine learning models have been utilized to investigate track defects on multi-dimensional parameters, and a plethora of factors affecting track quality have been identified [12]; however, little is known about the overall influence of data point similarities and dissimilarities on the prediction accuracy. According to Moreno-Torres et al., the disparities obtained due to the biases of the data affect the reliability of the data structure, and thus the classifier (machine learning method) accuracy is affected [13].

The study described herein attempted to investigate the effect of inherent track parameter disparities in the distribution domain. The parameters that were observed to have large variance were identified and corrected before the dimension reduction techniques were applied. The results show that while the machine learning methods may have looked sufficient for the track geometry problems, their accuracy tended to deteriorate or become less efficient as we continued to feed our model with different geometry data. The overall implementations revealed that the covariate shift implementation should not be downplayed where a slight change in the

[1]Corresponding author.

Contributed by the Computers and Information Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received March 12, 2021; final manuscript received June 22, 2021; published online July 14, 2021. Assoc. Editor: Ehsan T. Esfahani.

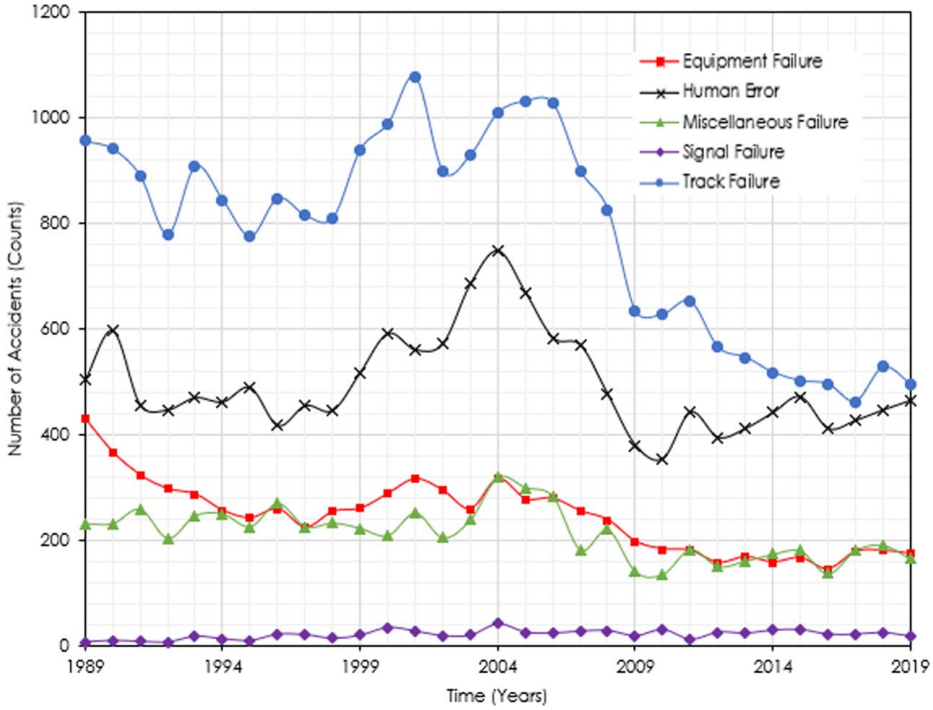

Fig. 1 Federal Railroad Administration rail accident distribution by components

parameters could drastically affect the model accuracy. The rest of the paper is described as follows. Section 2 discusses the track quality index (TQI) and associated parameters. In Sec. 3 of the paper, the authors introduce the adapted framework and the reduction techniques. The penultimate section introduces the data and its exploratory analysis, while the final section presents the conclusion and discussion.

2 Track Quality Index

While the continuous loading of the train system unarguably excites the track components, causing the degradation of the track geometry, track anomalies can be detected when the factors contributing to the accident are understood. Due to the complexity of the infrastructure, rail accident may be challenging to eliminate. Instead, track conditions may be improved with compliance with safety procedures [14]. This is because a quality track system is a requirement for safe rail operations [15]. Works of literature have shown that the majority of rail accidents in the United States can be described with the statistics shown in Fig. 1. However, in-depth accident distributions due to track geometry irregularities are less explored [16]. According to Higgins and Liu, train accidents could occur as a result of track geometry defect (geo-defect) or structural defects such as fasteners, sleepers, and clips [17]. Ivan Gallo et al. [18] reported that 34% of the recorded train accidents in 2009 were caused by track defects, resulting in a total damage of $108.7 million. Similarly, in 2012 over $102.9 million was spent on track restoration due to 33% of train accidents resulting from track defects [19]. The track manager needs to understand the interoperability of the track components to avoid unrecoverable train accidents. Thus, the TQI summarizes the deviation of the track geometry parameters from the threshold standards.

The TQI is mainly represented by the track geometry parameters, including the profile, alignment, warp, crosslevel, superelevation, and gauge [1], and it serves as a track fitness check for safe rail operations [12]. The two broad classes of TQI are single-track indices and combined indices. In the former model of evaluation, each track parameter is considered per unit length to carefully assess the signature and affirm the exact location of distress or offshoot. A typical FRA TQI is shown in Fig. 2, which is considered for this study; usually, a 200 ft track length measurement is used to avoid hypersensitivity of parameters. Figure 2 layout is distinctively defined in order to understand the geometry operations. In the combined TQI, the parameters are treated as a single entity, and their threshold levels are used to give the overall TQI. In general, the study of TQI is important because it enables early detection of critical states of the infrastructure. To date, different TQI exist due to the modification of track parameters to fit the country's safety thresholds. Some of the indexes are Canadian TQI, Poland TQI, FRA TQI, and Netherland TQI.

Table 1 Geometry safety thresholds: Class 4 track

	Preventive Class 4 safety thresholds (20 m)				
Alignment	Profile	Warp	Crosslevel	Gauge	Superelevation
32	51	44	32	1461	85

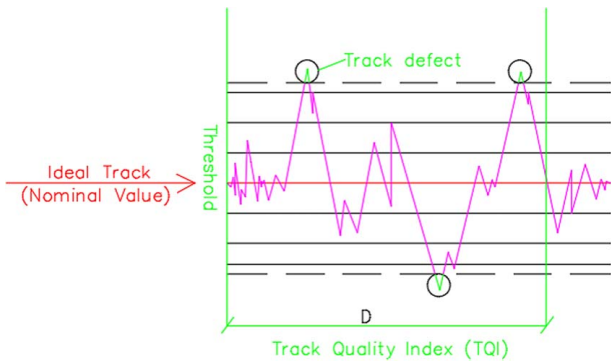

Fig. 2 A typical track geometry index of a railway system

In Poland, the TQI utilizes the synthetic track coefficient (J) to evaluate the track geometry condition based on the standard deviations (SDs) of different parameters. The index is expressed in the form:

$$J = \frac{S_z + S_y + S_w + 0.5 * S_e}{3.5} \tag{1}$$

where J represents the track quality index. The standard deviations of the geometry parameters, i.e., profile alignment, twist, and gauge, are denoted by S_z, S_y, S_w, S_e, respectively.

In the United States, the common track performance evaluator is the roughness index established by Amtrak. The index can be calculated by the average of squared differential geometry deviations over a chord length of 20 m as follows [20]:

$$r^2 = \frac{1}{n}\sum_{i=1}^{n-1}(G_{dev\ t+!} - G_{dev\ t})^2 \tag{2}$$

In the equation above, r^2 represents the track roughness value, n is the number of measurements, $G_{dev\ t+!}$ and $G_{dev\ t}$ represent the amount of gauge deviation for two consecutive years. A point of interest is that different indices utilize different chord lengths to evaluate the rail quality index. While the Polish TQI considers a chord length of 10 m, the Amtrak roughness index utilizes the 20 m chord length. Other acceptable quality indices can be found in the papers cited here [13,14].

The Canadian track quality index (CTQ) relies on a few parameters to ascertain the condition of the track systems. These are longitudinal level, alignment, gauge, and crosslevel. The TQI is obtained by averaging the results

$$TQI_i = 1000 - C * \sigma_i^2 \tag{3}$$

$$TQI = \frac{\sum_{i=1}^{n} TQI_i}{n} \tag{4}$$

For Netherlands TQI, the determining factors are the track segment and the standard deviation (SD) of the track parameters. Therefore, the sections that require maintenance can then be determined using the formula.

$$N = 10 * 0.675^{\frac{\sigma_i}{\sigma_i}^{80}} \tag{5}$$

N represents the index for an individual track geometry parameter.

It is important to note that while several indexes exist, the authors considered the Amtrack FRA TQI for this study.

2.1 Knowledge Gap. Track geo-defects and track are vital prerequisites for safe railroad operations. However, investigations revealed that even when all track geo-defects parameters are observed to conform within the threshold's standards, track accident still occurs. Due to the versatility and complexity of data structure, track managers decide the maintenance scheme that best describes track condition. Techniques that have existed are the statistical and stochastics method. Wang et al. [21] apply a statistical model to quantify the causes of the train accident in the United States based on the historical factors established by the FRA. The outcome can be used in the risk assessment of freight rail or general hazardous transportation across the North American corridors [21].

While the technology is being improved continuously, their capacity to store information also increases, leading to huge geometry data. The burden to analyze the present condition of the track system lies with the track engineer. In the past, both statistical and stochastic methods are helpful. However, the recent size of track data has forced analyst to machine learning methods. Lasisi and Attoh-Okine [22] apply the principal component analysis (PCA) to determine the track quality index of multi-dimensional geometry data. The study shows the possibility of transforming huge track geometry data into low sample space without compromising the geometry information [22].

Even though most track issues leading to rail accidents are reported to have come from continuous track data. Track engineers have also included image data analysis as bait for maintenance measurement.

One of the shortcomings identified with the machine learning methods is the low performance with big continuous data. Investigations have shown that depending on the data structure. Some problems are never entirely solved by the ML techniques, high-dimensional data, stream data, and covariate data. The reason is that the techniques are train and test on the same data. When they are exposed to new data, their adaptability diminishes. In order to increase the adaptability and accuracy of the machine learning methods, the concept of data shift should be understood.

Data shift problem has been explained in many fields with real-life problems except railroad. Whenever a shift occurs in data distribution, the accuracy of the solving techniques is usually affected. In railway engineering, considering a track data of 30 years, it will be difficult to isolate the likelihood of train accident for a season since rail track response to seasonal temperature. Thus, track data exploration for possible deterioration should be done with a consideration of external influence. Similarly, the clarity of the geometry vehicle's image during maintenance depends on the intensity of the light projected on the rail. Therefore, an image generated in the day is expected to differ from the night. According to Hajizadeh et al., 12,000 track defect images, including clip, rail joint, grinding mark, etc., were collected during a maintenance operation [23].

In this study, the authors consider the continuous track geometry data to verify possible shift before applying the desired machine learning. The track data will initially be reduced with the use of dimension reduction techniques. After that, covariate shift is examined. The purpose of this operation is that when ML techniques are applied without shift evaluation, there is a possibility of the same problem occurring in future.

3 Covariate Shift Optimization

The evolving nature of rail accident in North American demands continuous monitoring of track geometry data. Recent findings show that it is possible to have less accurate results using machine learning techniques due to the instability of data structure [24]. Selection bias and a nonstationary environment are the reason for the shift. Not handing the dataset shift in railway applications creates an overfitted model on training samples, hence unreliable model predictions. It is important to coordinate the distances between the data points. Therefore, dimension reduction techniques suffice.

There is a plethora of research on covariate shift adaptation where the main interest is to estimate the density ratio and improve the prediction in the target domain [25]. Some of the traditional methods fail to estimate the weight, especially in a high-dimensional space, except with robust classical algorithms [16–28]. According to the custom of data sampling, train and test data should be drawn from the same distribution. However, the circumstance which violates this assumption is known as covariate shift. Often, the problem of shift stems from the nonstationarity of the environment and selection bias, which are the primary reasons for its peculiarity across study fields. We address track defect classification problems for which the geometry features are dependent on the entire track parameter distribution that is allowed to differ arbitrarily from the target distribution. In railroad transportation, the evaluation of a covariate shift is never considered, while its influence could drastically affect machine learning performance [29].

Covariate shift refers to the change in the distribution of the input variables present in the train-test data shift, and it is widely discussed because of its canonical importance. It is the most common type of shift, and it is now gaining more attention as nearly every real-world dataset suffers from this problem. Data shifts have been classified into three major categories: covariate shift, prior probability shift, and concept shift [13–30].

In a given data distribution, drift is said to occur when the joint distribution changes over time from i to j:

$$P_i(X, Y) \neq P_j(X, Y) \tag{6}$$

On the other hand, a covariate shift will occur when

$$P_{y|x} = Q_{y|x} \text{ and } Q_x \neq P_x \tag{7}$$

When dealing with track geometry data, the input variables $\{x_1, x_2, x_3, \ldots\ldots x_n\}$ are referred to as covariates, and each feature represents one of the track geometry parameters. Inconsistent handling of the covariate data could result in overfitting of the training samples, leading to unreliable model performance. The objective function will then be the minimization of train-test divergence. In Eq. (4), when $P_x < Q_x$ and the distribution has density functions p_x and q_x, the standard approach is to replace the optimization error with the estimated density ratio:

$$Wx = p_x(x)/q_x(x) \tag{8}$$

Furthermore, the associated risk with track geometry samples can be idealized using Eqs. (9)–(12)

$$\underbrace{\textit{Track geometry data}}_{\textit{minimized train data}} : Q(x, y) = q(x)p(y|x) \tag{9}$$

$$\int dx p(x) \int dy p(y|x) l(f(x, w), y) \tag{10}$$

$$\underbrace{\textit{Track geometry data}}_{\textit{minimized train data}} : P(x, y) = p(x)q(y|x) \tag{11}$$

$$\int dx\, q(x) \int dy\, p(y|x) l(f(x, w), y) \tag{12}$$

In Eq. (8), the training P and hidden test Q can be modified to reduce the error using weights average. The weight serves in the place of eliminating the insufficient contributing data before the reduction process. The algorithm takes on the track geometry parameters: Twist L, Twist R, Alignment_10m_R, Alignment_10m_L, Alignment_20m_R, Alignment_20m_L, Profile_10m_R, Profile_10m_L, Profile_20m_R, Profile_20m_L, Cant R, Cant L, Superelevation, Guage, to predict the deterioration factor (y), which is dependent on the geometry parameters. For each iterative step, the loss function (l) in Eq. (10) measures the risk associated (f) with the classification error. With Eqs. (9) and (10), the task of minimizing the divergence in the training data is achieved. A similar operation is performed on the test data using Eqs. (11) and (12).

The shortfall in estimating the density ratio (w) in the expression above is the inadaptability to high-dimensional sample space. A more robust method is required to arrive at a low-dimensional space. Researchers have applied several methods to correct the shift. Reddi and Smola (2014) proposed a regularization method that controls the stability when correcting for covariate shift. Also, Wang and Rudin [31] explored the idea of dimensionality reduction to keep relevant information for posterior regression. The authors of this paper deployed unsupervised methods, PCA, and t-distributed stochastic neighbor embedding (TSNE) to reduce the track parameters' dimensionality. Section 3.1 introduces the hybrid techniques in detail.

3.1 Dimension Reduction Technique. Principal component analysis and TSNE are two popular feature extraction techniques. PCA is similar to TSNE except that the crowding problems with high-dimensional data are optimized. TSNE, on the other hand, uses the t-distribution to compute the similarity between points in a low-dimensional space; hyperparameters like perplexity, number of iterations, learning rate, and momentum are considered. These techniques have been applied to a railroad problem, but a shift was never investigated.

3.1.1 Principal Component Analysis. Principal component analysis is a foundational technique in machine learning, and it is agreed to be a fundamental algorithm for dimension reduction [32]. It has been applied to many fields, such as medicine, engineering, and science [33,34]. In railroad research, the PCA analysis serves as an important tool to better understand the distribution of track geometry parameters. The sample space is represented as $\varepsilon \mathrm{T}^{mxn}$ where m and n denote the track parameter and the inspection dates, respectively. Intuitively, each geometry parameter, such as gauge, superelevation, twist, or cant, is drawn from the random vector space X. As such, $X_1 \ldots . X_n\ \varepsilon \mathrm{T}^{mxn}$. PCA projects the track parameters "n" of the track data "X" onto a new orthogonal space, such that the new axes assume the directions of the largest n variance in the track data. A common problem with multiclass distribution is the uneven distribution of the individual features, affecting the data's overall performance. To cover for the inherent anomalies, each parameter is assigned a relative objective weight, "w," such that the parameters are now represented as $wX_1 + wX_2 \ldots\ldots wX_n$. Interestingly, since the parameters have the maximum variance, the representation of their magnitude in low-dimensional space gives its real value without compromising the individual attributes.

Furthermore, we represent the eigenvector corresponding to the largest eigenvalue (λ) of $X^T X$:

$$W = \underset{\|w\|_{x-1}}{\operatorname{argmax}} \sum_{i=1}^{n} (x_i^T w)^2 \tag{13}$$

3.1.2 t-Distributed Stochastic Neighbor Embedding. A more robust method for solving the dimensionality problem is using TSNE.

Given a track geometry data $N = \{x_1, x_2, \ldots., x_n\}$, TSNE computes a unique probability $P_{j|i}$ using distinct geometry observations x_i and x_j. Mathematically, the probability $P_{j|i}$ can be defined as

$$P_{j|i} = P_{i|j} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{14}$$

It is expected that the track geometry inspection of two distinct points can be represented as x_i and x_j. The collective data points follow a Gaussian distribution with the probability of point similarity given as $P_{j|i}$. However, a daunting situation occurs when the expected distinct distance is zero; the difference between x_i and x_j becomes zero. A way out is to adapt TSNE, which hypothetically creates a variance within the space.

In order to create a huge variance within the closely packed data, TSNE measures the similarities by learning a d-dimensional array $\{y_1, y_2, \ldots., y_n\}$

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|x_i - x_k\|^2)^{-1}} \tag{15}$$

We can explore the dissimilarities in a low-dimensional vector space using the Kullback-Leibler divergence [35]

$$KL(P|Q) = \sum_{K \neq 1} P_{i|j} \log \frac{p_{ij}}{q_{ij}} \tag{16}$$

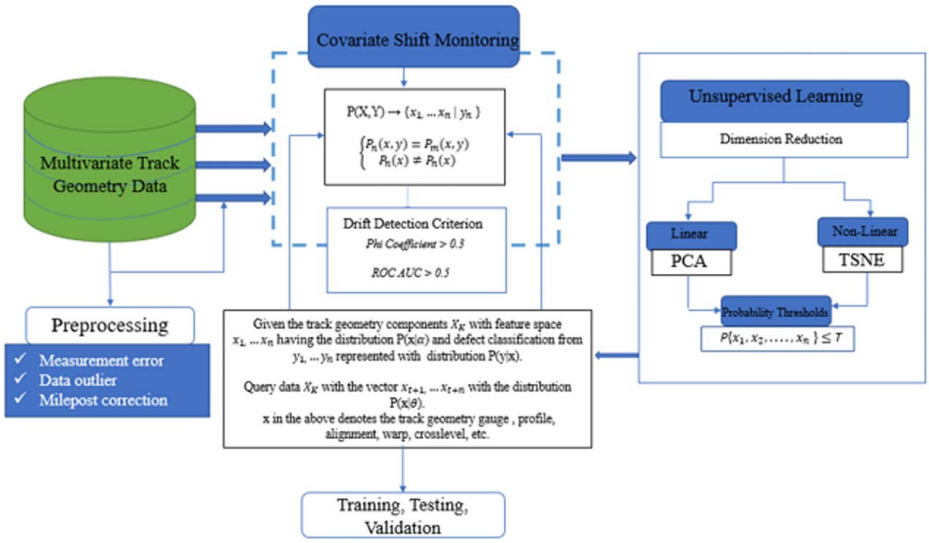

Fig. 3 Proposed covariate optimization framework

Algorithm for TSNE

Track geometry data $P = \{x_1, x_2, \ldots, x_n\}$
Cost function parameters: perplexity Perp
optimization parameters: number of iterations T, learning rate $\boldsymbol{\eta}$, momentum (t)
Result: low-dimensional data representation $\Upsilon^{(T)} = \{y_1, y_2, \ldots, y_n\}$
Begin
Compute pairwise probability $P_{j|i}$ with perplexity Perp
Set $p_{ij} = \frac{P_{j|i} + P_{i|j}}{2n}$
Sample initial solution $\Upsilon^{(0)} = \{y_1, y_2, \ldots, y_n\}$ from $\aleph(0, 10^{-4}\mathrm{I})$
For $t = 1$ to T do the following:
Compute low-dimensional probability q_{ij}
Compute gradient $\frac{\delta C}{\delta \Upsilon}$
Set $\Upsilon^{(T)} = \Upsilon^{(t-1)} + \boldsymbol{\eta} \frac{\delta C}{\delta \Upsilon} + (t)(\Upsilon^{(t-1)} - \Upsilon^{(t-2)})$
End

4 Methodology

This study's track geometry data represented both track inspection and maintenance data for a track Class 4 North American railroad. Overall inspection data showed 0.3% defects on the total track length of 82 km. The entire segment study revealed over 50 parameters and three hundred thousand observations. However, we considered 15 parameters for this research. This is because not all the parameters describe the condition of the track system. To carefully analyze the complex geometry data against inherent feature dissimilarities, the authors investigated the covariate shift and afterward corrected the divergence using the proposed framework in Fig. 3. The data are processed with a 16 GB Intel Core i7, 4.7 GHz. The multistep operations leading to robust predictive analysis involves data cleaning and dimension reduction. The process is repeated until a covariate shift is detected and removed. The shift operator thresholds are the phi coefficient ($\boldsymbol{\psi}$) and operating receiver characteristics (ROC). The shift is said to occur when the phi and ROC-AUC values are above 0.05 and 0.5, respectively.

5 Exploratory Data Analysis

In this section, we explored the track geometry components for a possible shift. The authors adopted two-way techniques to investigate the level of divergence. The first step dealt with visualization, as shown in Fig. 3, because we believed that some other parameters might not be obvious to visualize except with a metric thresholding algorithm. The visualization process is then preceded by analytical divergence evaluation to detect the diverging track parameters that cannot be captured virtually. Table 2 shows the captured track geometry parameters that significantly diverge from the threshold. While six of the components show no divergence, nine components significantly diverge. The divergence of the cant parameter is shown in Fig. 4. It means that the diverging components are the parameters influenced by external factors and should be iterated until no parameter diverges. The authors establish that careful consideration for these steps will ensure the accurate implementation of the machine learning techniques. To further investigate the extent of the divergence of the track geometry distribution, we subjected the train-test data to continuous metric thresholding (phi coefficient and ROC AUC value). The algorithm detected the covariate shift with different weight estimations. The iteration of the covariate shift detection continued until the required threshold was met. Thus, no covariate shift will be recorded at that threshold. This operation was conducted so that the new track geometry data and the train data can be seen as having a similar distribution. We initially codified the combined train and test distribution with a binary classifier and then predicted the test data's probability in the distribution.

Furthermore, we assigned a label to both train and test for the classifier. The level of the divergence showed the extent of the dissimilarities in the distribution. The predicted data in this study are three parameters: no defect, profile, alignment, and superelevation.

Table 2 Detected track geometry parameters with divergence values

Track parameters	Divergence
Superelevation	0.93
Twist_L	0.97
Twist_R	0.97
Alignment_10m_R	0.73
Alignment_20m_R	0.79
Profile_10m_R	0.70
Profile_20m_R	0.77
Profile_20m_L	0.81
Cant_R	0.96

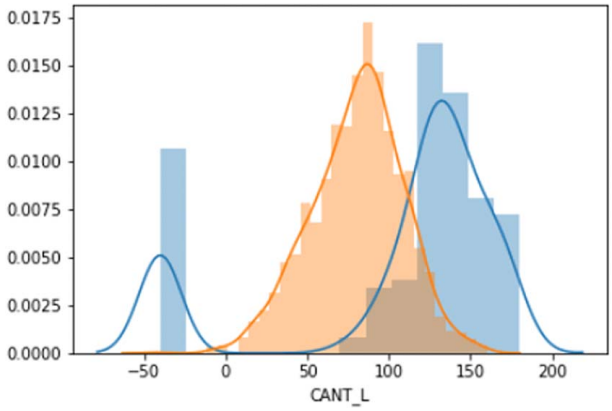

Fig. 4 Left Cant distribution with metric algorithm thresholding

These parameters are predicted because they have thresholds defined by FRA with which the condition of a track geometry can be concluded.

Covariate Shift Algorithm

Input X_{train} *and* X_{new}
For i in X_{train}, *assign YES if covariate shif exists, NO otherwise*
Assign 0 *to the prediction label and* 1 *to the new label*
Combine track geometry parameter X_{test} *and* X_{new} *with their prediction labels* Y_t *and* Y_{new}
Establish a distinct dataset $(X, Y)_T$
Perform a test on $(X, Y)_T \rightarrow (X, Y)^{\sim}_{train}$ *and* $(X, Y)^{\sim}_{test}$
Investigate a binary classifier $\hat{\kappa}$ *on* $(X, Y)^{\sim}_{train}$ *and* $(X, Y)^{\sim}_{test}$ *to detect shift*
Compute a shift value using the expected label from line 3 *and predicted value from line* 5
If $\psi > 0.2$ *and* $F_1 > 0.7$ *return YES*
else NO
End

6 Results and Discussion

It is essential to adequately verify any data distribution before the application of machine learning. This will guide the analyst against anomaly that could negatively influence the accuracy of the technique applied. In this study, we considered the implications of covariate shift evaluation of track geometry parameters. The raw geometry data contained several parameters where only a few of the parameters describe the condition of the track system. Thus, a reduction technique was considered. Prior to the implementation of the reduction techniques, the track geometry parameters were identified and resolved, as described in Tables 2 and 3. Each time we applied the techniques, we varied the predictors, i.e., alignment, profile, and no defect. The performance of the techniques differed because of the techniques of iterating parameters. This decision gave rise to each techniques being utilized with respective parameters. The track defects study with and without reduction techniques are presented in Table 4. Figures 5 and 6 show the visual representation of the predictive parameters in space with the PCA and TSNE techniques. From the figures, we observed that alignment is more dominated. It means that the clustered parameters have extremely minimal distances. Therefore, a shift in such instances can be detected with PCA.

A similar classification could be seen with profile, except that the TSNE detected the defect well enough compared with the PCA. A point of interest to the authors is the sudden drop in the accuracy of TSNE for profile detection. We found that the prediction can be enhanced with parameter tuning, such as perplexity. The perplexity creates a clearer topology in a distribution where the size, distance, and shape clustered. We evaluated the predictor with perplexity 1, 10, 25, and 50 to ascertain a better prediction (Fig. 7).

Table 3 Covariate shift detection of track geometry data

ψ Coefficient	ROC AUC	Covariate shift
0.72	0.9	Yes
0.61	0.87	Yes
0.55	0.82	Yes
0.43	0.77	Yes
0.16	0.5	No
0.96	0	Yes

Additionally, higher perplexity supports the redistribution of data points in space. The purpose of the perplexity tuning is to confirm information not captures by the TSNE holistically. In general, we observed that dimension reduction successfully is applied to multi-dimensional track geometry problems with consideration for shift detection to achieve better performance from the methods.

The evaluation of the track geometry with the dominant defects (profile and alignment) and without defect is presented in Table 4. The evaluators considered are two reduction techniques (PCA and TSNE) and linear model (parametrized geometry data). The accuracy result shows that the dimension reduction techniques are useful to analyze track geometry data with or without a likelihood of covariate shift. The accuracy is defined using the formula in Eq. (17)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$FNR = \frac{FN}{TP + FN} \tag{18}$$

$$FPR = \frac{FP}{TN + FP} \tag{19}$$

The True Positive (TP) shows the number of defects correctly assigned to the positive class, True Negative (TN) shows the number of defects correctly assigned to the negative class, False Positive (FP) denotes the number of defects assigned by the model to the positive class, which in reality belong to the negative class. False Negative (FN) denotes the number of defects assigned by the model to the negative class, which in reality belong to the positive class.

Table 4 The prediction accuracy of each technique for the track geometry defects

No defect			
Predictor	TPR	FPR	Accuracy
Unparameterized geometry data	0.9546	0.003	0.9976
TSNE-3D components	0.9135	0.067	0.9333
PCA 3	0.9135	0.067	0.9333
Profile			
Predictor	TPR	FPR	Accuracy
Unparameterized geometry data	0.9626	0.0017	0.9989
TSNE-3D components	0.3582	0.4736	0.4333
PCA 3	0.8765	0.1053	0.9000
Alignment			
Predictor	TPR	FPR	Accuracy
Unparameterized geometry data	0.9743	0.0015	0.9986
TSNE-3D components	0.8116	0.1772	0.8228
PCA 3	0.7972	0.2000	0.8000

TPR—True Positive Rate.
FPR—False Positive Rate.

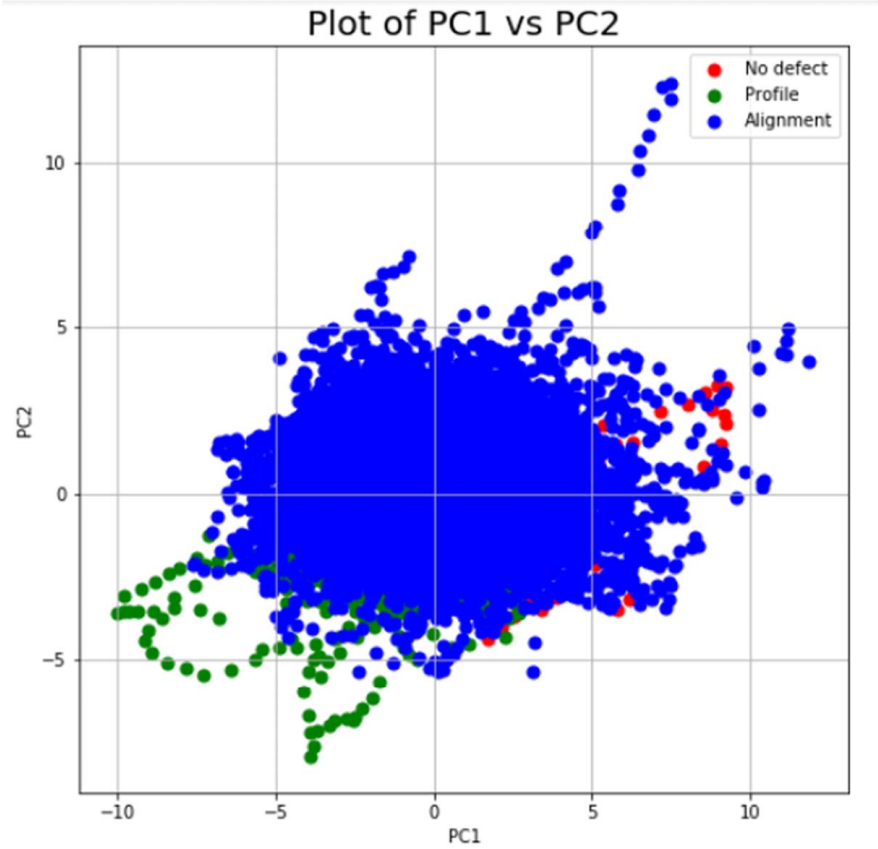

Fig. 5 Principal component analysis of the track geometry defects

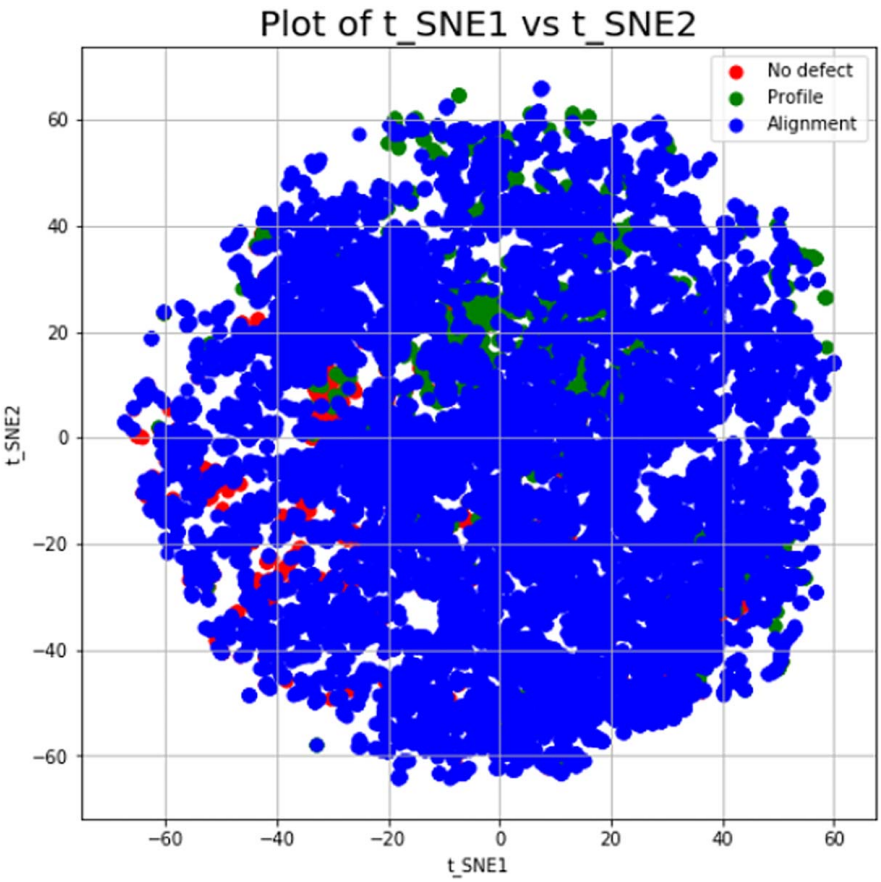

Fig. 6 TSNE analysis of the track geometry defects

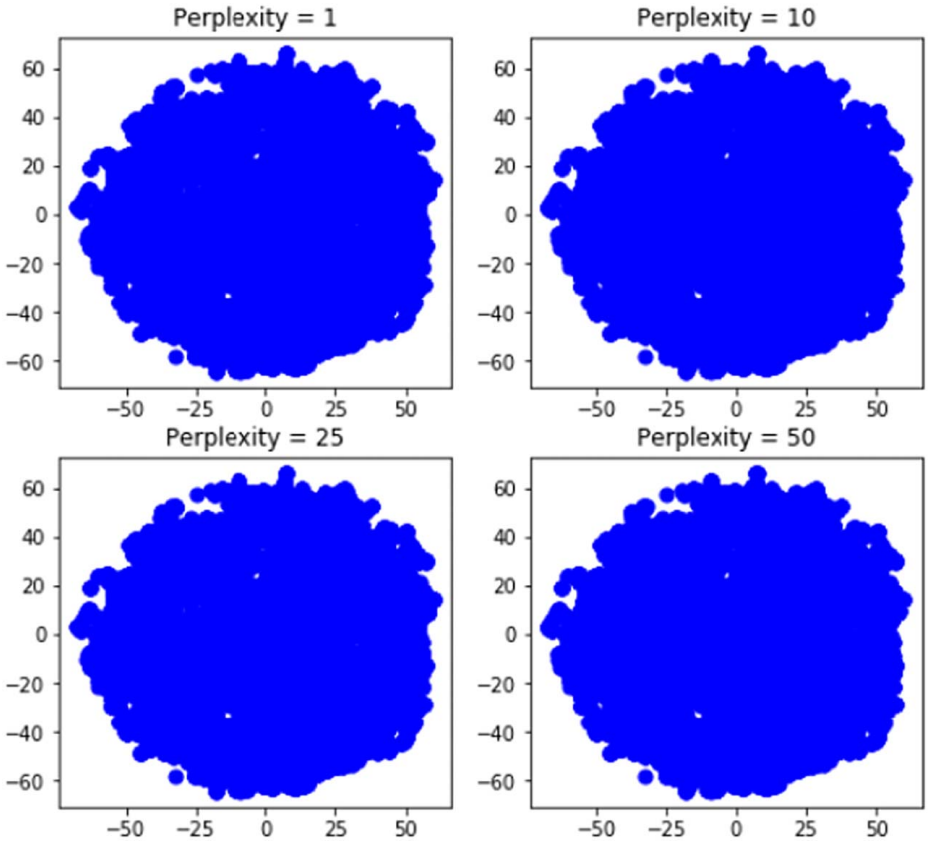

Fig. 7 TSNE analysis of the track geometry defect with varying perplexity values

7 Conclusion

One of the major challenges that affect the adaptability of vanilla machine learning to track geometry problems is the theory of "feature assumption." Here, the applied techniques are modelled to behave like all data distributions are identical and independently distributed. However, real-life problems tend to violate these assumptions. The field of machine learning has long been solving the issues of the railroad by unravelling the intricacies of complex data structures. Anecdotally, the application of covariate shift adaptation to railroad problems is relatively new, and we hope to explore it extensively. This technique is relevant in the sense that when data are trained to understand a certain trend, then detecting the trend of a new distribution could affect the performance of the machine learning technique.

This study addresses the issues of covariate shift detection optimization in track geometry problems. The track data was first subjected to covariate shift evaluation to curb the technically imbalanced distribution that could affect the accuracy of the machine learning techniques. Afterward, dimension reduction techniques were applied to reproduce the shift-cleared geometry data in a low-dimensional space. The results show that unique precision can be made without fear of future model degradation if the shift is adequately explored. Additionally, reduction techniques are also valuable for working with complex geometry data. Future work will consider robust covariate shift minimization techniques that would separate closely packed track data points, which are usually problematic for machine learning methods. Hence, improved track geometry defect predictions would be established.

Acknowledgment

Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of the sponsor.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Bakhtiary, A., Zakeri, J. A., and Mohammadzadeh, S., 2021, "An Opportunistic Preventive Maintenance Policy for Tamping Scheduling of Railway Tracks," Int. J. Rail Transp., **9**(1), pp. 1–22.
[2] Zhu, T., Xiao, S., Lei, C., Wang, X., Zhang, J., Yang, B., Yang, G., and Li, Y., 2020, "Rail Vehicle Crashworthiness Based on Collision Energy Management: An Overview," Int. J. Rail Transp., **9**(2), pp. 101–131.
[3] McAleer Law, 2021, "Train Accident Statistics," 2020, https://www.mcaleerlaw.com/train-accident-statistics.html, Accessed February 10, 2021.
[4] Kang, Y., Iranitalab, A., and Khattak, A., 2019, "Modeling Railroad Trespassing Crash Frequency Using a Mixed-Effects Negative Binomial Model," Int. J. Rail Transp., **7**(3), pp. 208–218.
[5] FRA, 2020, "Train Fatalities, Injuries, and Accidents by Type of Accident," National Transportation Statistics, https://www.bts.gov/content/train-fatalities-injuries-and-accidents-type-accidenta, Accessed February 8, 2021.
[6] Mazareanu, E., 2021, "United States—Rail Accidents and Incidents 2013–2019 Published by E. Mazareanu, Mar 27, 2020. This Statistic Represents the Number of Rail Accidents and Incidents in the United States From 2013 Through 2019. In 2019, the United States Registered 937 Rail," 2020, https://www.statista.com/statistics/204569/rail-accidents-in-the-us/.
[7] FRA.Gov, 2020, "Highway Rail Accidents," https://catalog.data.gov/dataset/highway-rail-accidents, Accessed February 8, 2021.
[8] Ambros, J., Perůtka, J., Skládaný, P., and Tučka, P., 2020, "Enhancing the Insight Into Czech Railway Level Crossings' Safety Performance," Int. J. Rail Transp., **8**(1), pp. 99–108.
[9] Wu, S. C., Xu, Z. W., Liu, Y. X., Kang, G. Z., and Zhang, Z. X., 2018, "On the Residual Life Assessment of High-Speed Railway Axles Due to Induction Hardening," Int. J. Rail Transp., **6**(4), pp. 218–232.

[10] Lasisi, A., Merheb, A., Zarembski, A., and Attoh-Okine, N., 2019, "Rail Track Quality and T-Stochastic Neighbor Embedding for Hybrid Track Index," 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, pp. 1470–1477.
[11] Lidén, T., 2015, "Railway Infrastructure Maintenance—A Survey of Planning Problems and Conducted Research," Transp. Res. Procedia, **10**(7), pp. 574–583.
[12] Lasisi, A., and Attoh-Okine, N., 2019, "Machine Learning Ensembles and Rail Defects Prediction: Multilayer Stacking Methodology," ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng., **5**(4), p. 04019016.
[13] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F., 2012, "A Unifying View on Dataset Shift in Classification," Pattern Recognit., **45**(1), pp. 521–530.
[14] Sun, Y. Q., 2018, "Mitigating Train Derailments Due to Sharp Curve and Overspeed," Front. Mech. Eng., **4**(8), pp. 1–12.
[15] Offenbacher, S., Neuhold, J., Veit, P., and Landgraf, M., 2020, "Analyzing Major Track Quality Indices and Introducing a Universally Applicable TQI," Appl. Sci., **10**(23), pp. 1–17.
[16] He, Q., Li, H., Bhattacharjya, D., Parikh, D. P., and Hampapur, A., 2015, "Track Geometry Defect Rectification Based on Track Deterioration Modelling and Derailment Risk Assessment," J. Oper. Res. Soc., **66**(3), pp. 392–404.
[17] Higgins, C., and Liu, X., 2018, "Modeling of Track Geometry Degradation and Decisions on Safety and Maintenance: A Literature Review and Possible Future Research Directions," Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit, **232**(5), pp. 1385–1397.
[18] Cárdenas-Gallo, I., Sarmiento, C. A., Morales, G. A., Bolivar, M. A., and Akhavan-Tabatabaei, R., 2017, "An Ensemble Classifier to Predict Track Geometry Degradation," Reliab. Eng. System Safety, **161**, pp. 53–60.
[19] He, Q., Li, H., Bhattacharjya, D., Parikh, D. P., and Hampapur, A., , 2015, "Track Geometry Defect Rectification Based on Track Deterioration Modelling and Derailment Risk Assessment," J. Oper. Res. Soc., 66(3), pp. 392–404.
[20] Falamarzi, A., Moridpour, S., and Nazem, M., 2021, "A Time-Based Track Quality Index: Melbourne Tram Case Study," Int. J. Rail Transp., **9**(1), pp. 23–38.
[21] Wang, B. Z., Barkan, C. P. L., and Rapik Saat, M., 2020, "Quantitative Analysis of Changes in Freight Train Derailment Causes and Rates," J. Transp. Eng. A: Syst., **146**(11), p. 04020127.
[22] Lasisi, A., and Attoh-Okine, N., 2018, "Principal Components Analysis and Track Quality Index: A Machine Learning Approach," Transp. Res. Part C Emerg. Technol., **91**(3), pp. 230–248.
[23] Hajizadeh, S., Núñez, A., and Tax, D. M. J., 2016, "Semi-Supervised Rail Defect Detection From Imbalanced Image Data," IFAC-PapersOnLine, **49**(3), pp. 78–83.
[24] Balogun, I., and Attoh-Okine, N., 2021, "Random Forest-Based Covariate Shift in Addressing Nonstationarity of Railway Track Data," ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng., **7**(3), p. 04021028.
[25] Polo, F. M., and Vicente, R., 2020, "Covariate Shift Adaptation in High-Dimensional and Divergent Distributions," arXiv, 1.
[26] Sharma, S., Cui, Y., He, Q., Mohammadi, R., and Li, Z., 2018, "Data-Driven Optimization of Railway Maintenance for Track Geometry," Transp. Res. Part C Emerg. Technol., **90**(9), pp. 34–58.
[27] Reddi, S. J., Poczos, B., and Smola, A., 2015, "Doubly Robust Covariate Shift Correction," AAAI Conference on Artificial Intelligence, Austin, TX, Jan. 25–30, pp. 1–14.
[28] Wang, S., McCormick, T. H., and Leek, J. T., 2020, "Methods for Correcting Inference Based on Outcomes Predicted by Machine Learning," Proc. Natl. Acad. Sci., **117**(48), pp. 30266–30275.
[29] Soleimanmeigouni, I., Ahmadi, A., and Kumar, U., 2018, "Track Geometry Degradation and Maintenance Modelling: A Review," Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit, **232**(1), pp. 73–102.
[30] Dharani, G., Nair, N. G., Satpathy, P., and Christopher, J., 2019, "Covariate Shift: A Review and Analysis on Classifiers," 2019 Global Conference for Advancement in Technology (GCAT) 2019, Bangalore, India, Oct. 18–20, pp. 1–6.
[31] Wang, F., and Rudin, C., 2017, "Extreme Dimension Reduction for Handling Covariate Shift," arXiv.
[32] Joshi, Y., 2021, "Applications of Principal Component Analysis (PCA)," https://iq.opengenus.org/applications-of-pca/, Accessed January 13, 2021.
[33] Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., and McLoone, S., 2013, "Principal Component Analysis on Spatial Data: An Overview," Ann. Assoc. Am. Geogr., **103**(1), pp. 106–128.
[34] Alken, P., Maute, A., Richmond, A. D., Vanhamäki, H., and Egbert, G. D., 2017, "An Application of Principal Component Analysis to the Interpretation of Ionospheric Current Systems," J. Geophys. Res.: Space Phys., **122**(5), pp. 5687–5708.
[35] García-Alonso, C. R., Pérez-Naranjo, L. M., and Fernández-Caballero, J. C., 2014, "Multiobjective Evolutionary Algorithms to Identify Highly Autocorrelated Areas: The Case of Spatial Distribution in Financially Compromised Farms," Ann. Oper. Res., **219**(1), pp. 187–202.