Article

A Hybrid Framework Combining Data-Driven and Catenary-Based Methods for Wide-Area Powerline Sag Estimation

Yunfa Wu 1, Bin Zhang 1,*, Anbo Meng 1, Yong-Hua Liu 1 and Chun-Yi Su 1,2

1 School of Automation, Guangdong University of Technology, Guangzhou 510006, China; wuyunfa@outlook.com (Y.W.); menganbo@vip.sina.com (A.M.); yonghua.liu@outlook.com (Y.-H.L.); cysu@scut.edu.cn (C.-Y.S.)
2 Department of Mechanical Engineering, Concordia University, 1455 de Maisonneuve Blvd. W., Montreal, QC H3G 1M8, Canada
* Correspondence: zhangbin_csu309@163.com

Abstract: This paper is concerned with the airborne-laser-data-based sag estimation for wide-area transmission lines. A systematic data processing framework is established for multi-source data collected from power lines, which is applicable to various operating conditions. Subsequently, a k-means-based clustering approach is employed to handle the spatial heterogeneity and sparsity of powerline corridor data after comprehensive performance comparisons. Furthermore, a hybrid model of the catenary and XGBoost (HMCX) method is proposed for sag estimation, which improves the accuracy of sag estimation by integrating the adaptability of catenary and the sparsity awareness of XGBoost. Finally, the effectiveness of HMCX is verified by using power data from 116 actual lines.

Keywords: sag estimation; hybrid model; ensemble learning; spatial data

1. Introduction

The clearance distance of overhead power lines under various working conditions must meet safety requirements to prevent accidents such as electric shock, discharge, and short circuits. The clearance distance is directly affected by the sag fluctuation; high-precision sag estimation is of great value for maintenance and expansion in power inspection and scheduling [1–6]. The sag varies dynamically with conductor temperature and horizontal stress [7]. The changes in operating conditions and fluctuations in the surrounding environment make it difficult to accurately estimate the sag of wide-area overhead power lines. With the advancement of intelligent inspection, the problem of accurate sag monitoring or estimation has attracted the attention of academia and industry again [8–11].

In the field of sag estimation, some researchers try to realize the on-line monitoring of sag by means of various measurement techniques and monitoring equipment. For example, Mahaj et al. [12] calculated the value of the sag by recording the GPS signals corresponding to the physical movement of the lines. Pan et al. [13] derived the sag directly through image recognition of the conductor employing an HD camera. Some researchers calculate the sag through sensors such as tension sensors [14], temperature sensors [15], fiber bragg grating sensors [16], and magnetic field sensor arrays [17]. Kopsidas et al. [18] developed a holistic method to calculate the sag and ampacity based on the mechanical and electrical parameters of the overall system. Among the above methods, the accuracy of the sag method based on tension and temperature is higher because they are variables directly related to sag variation. However, limited by the equipment cost, Such methods are only applicable to a small part of the line with equipment already installed rather than generalizing to wide-area transmission lines.

With the aid of advanced sensing techniques and intelligent algorithms, numerous researchers have sought to solve the problem of sag estimation for wide-area lines.
Du et al. [19] utilized the phasor measurement units (PMU) to estimate the average temperature of an entire line and then calculated the sag of the wide-area lines based on the intrinsic relationship between the resistance and temperature of the wire. While the PMU-based method makes it possible to estimate the arc sag for the entire line, such temperature-dependent sag estimation methods are subject to cumulative errors due to the continuous influence of micrometeorology around the line. For sag estimation of wide-area lines, methods that require the installation of additional devices or sensors are costly to the utility. In addition, Chen et al. [20] and Golinelli et al. [21] obtained the point cloud data of each line with laser scanning technology and derived the operation status of the transmission line (including sag) through steps of classifying, extracting and numerically analyzing.

The popularization of Airborne Laser Scanning (ALS) technology in power inspection makes it more convenient to obtain the corridor information on power lines [22–24]. ALS is also known as LiDAR (Light Detection and Ranging). It can easily obtain the point clouds of the line, and extract the sag and clearance distance to ground objects under scanning conditions. The application of ALS accumulates a large amount of multi-source corridor data. Limited by the scanning mode, the collected transmission line data is static, and belongs to spatial information [20,25]. The corridor data has spatial uncertainty in different locations [26]. Due to spatial size, regional influence, or spatial interaction of the LiDAR, corridor data have spatial dependence and spatial heterogeneity [27–30]. Nevertheless, with the development of machine learning, numerous algorithms with better robustness and generalization errors have spawned, such as eXtreme Gradient Boosting (XGBoost), which provides a better instrument for solving estimation problems [31–35].

LiDAR has been utilized in a variety of fields, including terrain mapping [36], mining measurements [37], forestry surveys [38], power inspections [39], disaster emergency responses [40], and smart cities [41]. In the power industry, LiDAR-based machine learning functions focus on point cloud classification and element detection and extraction, including power line and tower detection and extraction, power corridor vegetation management, power line scene classification and reconstruction, channel anomaly detection, and so on [42]. Several deep learning architectures for 3D point cloud classification, such as PointNet++ [43] and SparseCNN [44], have been developed in recent years [45]. References [46,47] classify ALS point clouds using a computer vision deep learning model with adaptive parameters for large scenes of outdoor ALS point clouds. As far as the authors know, further machine learning-based data mining, regression analysis based on LiDAR data is still in its early stages.

Motivated by the above-mentioned discussions, in this paper, a hybrid model of the catenary and XGBoost (HMCX) method is proposed to address the problem of sag estimation based on corridor data. The main contributions of this article are outlined fourfold as follows:

1. A systematic data processing framework for aircraft-based inspection is established to preprocess the multi-source heterogeneous corridor data.
2. A similarity clustering algorithm based on k-means is introduced to solve the problem of spatial heterogeneity and spatial dependence in the corridor data.
3. A novel HMCX method for sag estimation through utilizing corridor data is proposed, which combines the adaptability of catenary with the sparsity awareness of XGBoost.
4. The feasibility and effectiveness of HMCX are verified by using power data from 116 actual lines. The proposed HMCX method outperforms catenary, linear Regression, and Bayesian ridge regression involved in this study.

The remainder of the paper is structured as follows: An overview of the corridor database is presented in Section 2. The whole data analysis framework using corridor data and the proposed HMCX method for sag estimation are designed in Section 3. The explanatory experimental results include data analysis results, the performance of HMCX, and comparisons with other algorithms are summarized in Section 4. Eventually, the conclusions and some future work are discussed in Section 5.
Notations: All symbols used herein are in standard form unless otherwise indicated. Table 1 shows the symbols used in this paper and their meanings.

Table 1. Table of symbols.

| Symbol | Typical Meaning |
|--------|-----------------|
| a, b, c, α, β, γ | Scalars are lowercase |
| x, y, z | Column vectors are bold lowercase |
| A, B, C | Matrices are bold uppercase |
| x = (x₁, ..., xₙ), xᵢ ∈ R | n × 1 dimensional column vector |
| xᵀ = (x₁, ..., xₙ)ᵀ, xᵢ ∈ R | Transpose of a column vector or 1 × n dimensional row vector |
| X ∈ Rᵐˣⁿ or X = [xᵢⱼ]ᵐˣⁿ, xᵢⱼ ∈ R | m × n dimensional matrix |
| B = [b₁, b₂, b₃] | (Ordered) tuple |
| B = [b₁, b₂] | Matrix of column vectors stacked horizontally |
| B = [b₁, b₂, b₃] | Set of vectors (unordered) |
| B = [b₁, b₂, b₃] | Number of elements in the set, | |
| | | | |
| | | Cluster, where bᵢ is a vector in Bᵢ, Bᵢ is a set of vectors, and B is set of Bᵢ | |
| | | The set X includes m instances, each instance is represented by vector xᵢ of n attributes. | |
| | | Integers, positive integers and natural numbers, respectively | |
| | | Real numbers and positive real numbers, respectively | |
| | | n-dimensional vector space of real numbers | |

1 In the following context, X is also treated as [x₁, x₂, ..., xₙ] to support matrix operations, Mat(X) is used to matrix X for the distinction representation, which is not mathematically rigorous.

2. Description of Corridor Database

In order to obtain high-precision corridor information, electric power institutions collect point cloud models of transmission line corridors through laser scanners mounted on fixed-wing, helicopter, and multi-rotor UAVs. Limited by the length of a single operation of the aircraft, the lines are mostly in the form of short line segments. This paper collects the corridor data of 36 overhead transmission lines obtained from 116 ALS jobs in Guangdong Power Grid, and the year is from January 2018 to June 2020. The number of ALS operation records for 550 kV, 220 kV, and 110 kV is 76, 34, and 4 times, respectively. By using python for data processing and munging, we got the data of 30,945 phase lines of spans.

The flight platforms used in the ALS operation of the lines studied in this paper are mainly manned/unmanned helicopters and fixed-wing UAVs. A laser scanner is mounted on the flight platform, thereby realizing the aircraft-based inspection for power transmission lines. Figure 1 shows a picture of a helicopter-borne laser scanning inspection operation. The laser scanner models used in the studied lines are RIEGL VUX-1LR and RIEGL miniVUX-3UAV. The processing software of the collected line point cloud is LiDAR360.

After data extraction, cleaning, munging, transformation, matching, and load, a corridor database containing 30,945 phase lines of spans with 20 variables or features is generated, where the variables include the weather information, conductor parameters, terrain, point clouds, etc. The variables or constants of the database are shown in Table 2.

For ALS operation in the line corridor, the following information is recorded:

1. The ambient temperature and wind speed of the take-off site are recorded at the beginning of ALS operations.
2. The classified point cloud information is extracted from LiDAR, including the span length, height difference, sag value, and distance from the maximum sag point to the tower of each line span.
3. The conductor parameters, voltage, tower type, and service time are recorded in the ledger.

![Figure 1. Helicopter-borne laser scanning inspection operation.](image)

| Table 2. The selected features and their abbreviations. |
|--------------------------------------------------------|
| **Category** | **Item** | **Abbreviation** |
| Line information | Voltage level | VTG |
| | Span type | SPT |
| | Conductor type | CDT |
| | Terrain | TR |
| | Service time | ST |
| LiDAR data | Span length | SPL |
| | Height difference | HD |
| | Distance from the maximum sag point to the tower | DMT |
| | Maximum sag | MS |
| Take-off weather | Ambient temperature | AT |
| | Wind speed | WS |
| Conductor parameters | Elastic coefficient | ELC |
| | Breaking force | BRF |
| | Resistance per kilometer | RPK |
| | Diameter of wire | DW |
| | Linear expansion coefficient | LEC |
| | Weight per unit length | WPL |
| | Total cross sectional area | TCSA |
| | Steel core diameter | SCD |

3. Methodology

This section presents the construction process of the HMCX. First, a description of the data processing for corridor data is introduced, and a processing framework, including data extraction, data preprocessing, and data integration is proposed. Then, the catenary model of sag and its characteristics are described. Next, a similarity cluster analysis based on k-means is performed on the data with the catenary error. Finally, the HMCX is designed. The flow chart of this section is shown in Figure 2.
3.1. Data Processing

The data processing and analysis framework is established to solve the problem of multi-source heterogeneous data aggregation and analysis, as shown in Figure 3. The four phases of data extraction, data transformation, data integration, and database generation are mainly responsible for data processing.

Figure 3. The framework of data processing.
Various kinds of information are extracted from the data, including line information, LiDAR data, and environmental data. Table 2 lists the variables or parameters of 20 kinds that were extracted from each data source. The variables or parameters are transformed by four standard steps. The following are the specific steps:

1. Data consistency processing. It includes processing the format and content of multi-source data, unifying units and representations, and performing consistent processing to facilitate subsequent processing.

2. Data interpretation and transformation. The actual meaning of the parameters of the multi-source data is interpreted, and the unit is uniformly converted. Discrete variables of span type and terrain information are processed using one-hot encoding, the wire type and its parameters are matched to the dataset. The standardization is accomplished on continuous variables. The date feature is converted to the number of days the line has been put into operation; the tower coordinates at both ends of one span are converted to the Euclidean distance between them.

3. Missing values, duplicate values, and outliers handling use random forest regression to impute and fill features with fewer missing values and identify and eliminate duplicates and outliers in the dataset.

4. Feature analysis, selection, and reduction. In order to eliminate the influence of redundant variables, the feature selection based on gradient boosting decision tree (GBDT) is used to analyze the importance of features. Kernel principal component analysis (PCA) is used to determine whether the information between features is redundant and to reduce dimensionality.

5. Data integration is the process of using data from different sources to construct a unified view. Data from multiple sources can be linked through the mapping relationship of the common parameters. The corridor database is merged and updated by matching, redundant fusion, cooperative fusion, and complementary fusion.

According to the importance of features, the feature variables with low importance ranking are eliminated. After data processing, we obtained a dataset of 15 sag-related parameters. For N spans, the 15 sag-related parameters are described as follows:

- Span type (SPT) \( p = (p_1, \ldots, p_N) \) (\( p_i \in \{0, 1, 2, 3\} \)). The span type \( p_i \) includes double linear towers, double-sided tension towers, left single tension towers, and right single tension towers, which are represented as 0, 1, 2, and 3 by one-hot encoding, respectively.

- Conductor type (CDT) \( C = \{c_1, \ldots, c_N\} \) (\( c_i \in \mathbb{R}^{1 \times 7} \)), where \( c_i = (e_i, b_i, r_i, d_i, w_i, a_i) \times 7 \), \( (e_i, b_i, r_i, d_i, w_i, a_i) \) is one of the conductor types, and it is the parameters of the conductor of the \( i \)-th span. The parameters are elastic coefficient (ELC), breaking force (BRF), resistance per kilometer (RPK), diameter of wire (DW), linear expansion coefficient (LEC), weight per unit length (WPL), and total cross-sectional area (TCSA), respectively.

The respective vector representations of the conductor parameters in the dataset are as follows: ELC is \( e = (e_1, \ldots, e_N) \) (\( e_i \in \mathbb{R}^* \)), BRF is \( b = (b_1, \ldots, b_N) \) (\( b_i \in \mathbb{R}^* \)), RPK is \( r = (r_1, \ldots, r_N) \) (\( r_i \in \mathbb{R}^* \)), DW is \( d = (d_1, \ldots, d_N) \) (\( d_i \in \mathbb{R}^* \)), LEC is \( l = (l_1, \ldots, l_N) \) (\( l_i \in \mathbb{R}^* \)), WPL is \( w = (w_1, \ldots, w_N) \) (\( w_i \in \mathbb{R}^* \)), and TCSA is \( a = (a_1, \ldots, a_N) \) (\( a_i \in \mathbb{R}^* \)).

Terrain (TR) \( u = (u_1, \ldots, u_N) \) (\( u_i \in \{0, 1, 2, 3, 4, 5\} \)). The terrain mainly includes hills, mountains, plains, rivers, and farmland. We used one-hot encoding to number them 0, 1, 2, 3, 4, 5, respectively.

Service time (ST) \( s = (s_1, \ldots, s_N) \) (\( s_i \in \mathbb{R}^* \)). The service time is obtained by converting the date into the number of days the line has been put into operation.

Span length (SPL) \( l = (l_1, \ldots, l_N) \) (\( l_i \in \mathbb{R}^* \)). The span length \( l_i \) is obtained by converting the coordinates of the suspension points at both ends of a span into Euclidean distance.

Height difference (HD) \( h = (h_1, \ldots, h_N) \) (\( h_i \in \mathbb{R} \)). The height difference \( h_i \) is obtained by the difference between the heights of the suspension points at both ends.

Maximum sag (MS) \( f = (f_1, \ldots, f_N) \) (\( f_i \in \mathbb{R}^* \)). In ALS operations, point cloud quality problems such as ghost points, jump points, noise points, and missing caused by the jitter of the aircraft fuselage often occur. A differential threshold filtering method is utilized to smooth the point cloud.
Ambient temperature (AT) \( t = (t_1, \ldots, t_N) (t_i \in \mathbb{R}) \) and wind speed (WS) \( v = (v_1, \ldots, v_N) (v_i \in \mathbb{R}) \) are recorded in the take-off site at the beginning of ALS operations. Therefore, the \( i \)-th span data can be represented as a tuple \( P_i = (x_i, y_i) \), where \( x_i = (p_i, e_i, b_i, r_i, d_i, i, w_i, a_i, u_i, s_i, l_i, h_i, t_i, v_i) \), and \( y_i = f_i \), containing 15 sag-related variables in total. Finally, the corridor dataset can be represented as \( \mathcal{P} = \{P_i\}^N_{i=1} = \{P_1, \ldots, P_N\} \). From a practical point of view, its matrix form can be treated as:

\[
\text{Mat}(\mathcal{P}) := [p, e, b, r, d, i, w, a, u, s, l, h, t, v, f] \in \mathbb{R}^{N \times 15}
\]  

(1)

3.2. Catenary Model for Sag Calculation

This section introduces the catenary model of sag calculation and the calculation method of sag difference.

3.2.1. Catenary-Based Sag Calculation

The catenary model is frequently used to estimate line rough sag, because its physical parameters are clear and highly adaptable. The line between two suspension points is usually assumed to be a catenary line, which is assumed to be a chain of flexible cables without rigidity. The mechanism model of catenary is obtained from the force balance relationship. The point cloud of a span is shown in Figure 4 after classification.

![Figure 4. The catenary curve and sag of overhead transmission lines.](image)

The catenary model of the curve is:

\[
y = \frac{2\sigma_0}{\gamma} s h \frac{\gamma x}{2\sigma_0} - \frac{\gamma(x - 2a)}{2\sigma_0},
\]  

(2)

where:

\[
a = \frac{l}{2} - \frac{\sigma_0}{\gamma} \text{arcsh} \frac{h}{2\sigma_0} \left(\frac{1}{2} + \frac{1}{2}\right)
\]  

(3)

and \( \sigma_0 \) denotes horizontal stress, \( \gamma \) represents the load per unit length and cross-sectional area of the wire, \( x \) is the horizontal distance from the left suspension point.

The sag at any point \( x \) of the overhead line of the unequal height suspension points is:

\[
f_x = \frac{h}{l} x - y = \frac{h}{l} x - \frac{2\sigma_0}{\gamma} s h \frac{\gamma x}{2\sigma_0} - \frac{\gamma(x - 2a)}{2\sigma_0},
\]  

(4)

where \( l \) denotes the span length, \( h \) denotes the height difference between two suspension points.
The maximum sag appears at $df_x/dx = 0$, the maximum sag of the unequal height suspension point can be obtained as:

$$f_m = \frac{c_0}{\gamma} \frac{h}{l} (\text{arcsh} \frac{h}{L_{h=0}} - \text{arcsh} \frac{h}{\frac{2c_0}{\gamma} \text{sh} \frac{2h}{2c_0}})$$

$$+ \sqrt{1 + \left( \frac{h}{L_{h=0}} \right)^2 \frac{\gamma}{2c_0}} - \sqrt{1 + \left( \frac{h}{\gamma} \right)^2},$$

(5)

where $L_{h=0}$ is the length of the catenary in the span of the overhead line of the same height suspension point, namely:

$$L_{h=0} = \frac{2c_0}{\gamma} \text{sh} \frac{2h}{2c_0}.$$  

(6)

In practice, the above relationship is often used to calculate the sag of the line, which has clear physical information and is not complicated to implement.

The catenary model parameters are easy to determine using the $X_i$ data of the corridor dataset $\mathcal{P}$ in Section 3.1. Wind speed and conductor information are used to determine the comprehensive specific load. The horizontal stress is determined according to the safety factor and conductor parameters. Combining the span information in the $X_i$, the calculated sag value $\hat{f}_i$ of span $i$ can be obtained by substituting it into (5). Thus, the calculated sag of the catenary model for the corridor dataset can be represented as:

$$\mathcal{F} = \{\hat{f}_1, \ldots, \hat{f}_N\} (\hat{f}_i \in \mathbb{R}^*).$$  

(7)

3.2.2. Sag Difference between the Catenary and the Extracted Sag

The catenary model belongs to the white-box model, which is an accurate mathematical model based on the wire’s internal force mechanism. The catenary model does not account for the effect of drift in parameter values caused by aging, which brings great uncertainty to the calculation. Therefore, the catenary model has the obvious disadvantage of large calculation errors.

Based on $x_i$ of the corridor dataset $\mathcal{P}$ obtained in the previous subsection, it is easy to obtain the calculated sag value $\hat{f}_i$ of span $i$ by substituting it into (5). Therefore, the catenary model error between calculated sag $\hat{f}_i$ and the extracted sag $f_i$ from the LiDAR of the $i$-th span, namely, sag difference is:

$$\Delta f_i = f_i - \hat{f}_i.$$  

(8)

The sag differences from the catenary error corresponding to the corridor dataset is: $\delta f = \{\Delta f_1, \ldots, \Delta f_N\} (\Delta f_i \in \mathbb{R})$. Furthermore, the corridor dataset with errors is updated as:

$$\mathcal{P}_e = \{P_{e1}, \ldots, P_{eN}\} = \{x_i, \delta_i\}.$$  

(9)

where $x_i = (p_i, e_i, b_i, r_i, d_i, t_i, w_i, a_i, u_i, s_i, l_i, h_i, t_i, v_i)$, and $\delta_i = \Delta f_i$. Then the matrix forms are obtained:

$$x = (x_1, \ldots, x_N)$$

$$\delta = \delta f.$$  

(10)

The matrix form of all datasets considering catenary-based error, from a practical point of view, can be expressed as:

$$\text{Mat}(\mathcal{P}_e) := [X, \delta] = [p, e, b, r, d, t, w, a, u, s, l, h, t, v, \delta].$$  

(11)
3.3. k-Means-Based Similarity Clustering Considering Sag Difference

To reduce the impact of spatial sparsity and heterogeneity on estimation outcomes, we consider employing a clustering method to partition the dataset to improve similarity within one cluster. In order to choose an appropriate clustering method, we compared the effect of six clustering algorithms on corridor data clustering. Through the three performance indicators and the Kappa-based consistency test of the clustering results, we finally selected k-means and determined its initial value conditions. See Section 4.2 for more details.

For dataset $\mathcal{P}_e$ with catenary error, we employ the Euclidean distance scheme of k-means for clustering. The k-means algorithm divides a set of $N$ samples $X$ into $k$ disjoint clusters $C$, each cluster $C$ is described by the mean $\mu_j$ of the samples in the cluster (often called the centroid of the cluster). The k-means algorithm aims to choose centroids $\mu_j$ that minimise the inertia, or sum-of-squares criterion within cluster:

$$\sum_{i=0}^{n} min(||x_i - \mu_j||^2).$$

(12)

The cluster number $k$ needs to be determined in advance, and the positions of the $k$ initial centroids have a great impact on the clustering results and running time. $k$ is determined by comparing three performance indicators under a different numbers of clusters.

Finally, the corridor dataset with errors $\mathcal{P}_e = \{\mathcal{P}_{ei}\}_{i=1}^N$ is grouped into $k$ cluster divisions. The set of $k$ clusters is represented as:

$$\mathcal{C} = \{C_1, C_2, \ldots, C_k\} (C_i \in \mathcal{P}_e, k \in \mathbb{Z}^* \text{ and } k < N),$$

(13)

where $C_1 \cup C_2 \cup \cdots \cup C_k = \mathcal{P}_e$.

The dataset of cluster $C_i$ can be denoted as:

$$\mathcal{P}_{C_i} := \{\mathcal{P}_{ei}\}_{i=1}^n = \{\mathcal{P}_{ei}^{(1)}, \mathcal{P}_{ei}^{(2)}, \ldots, \mathcal{P}_{ei}^{(n)}\} (\mathcal{P}_{ei}^{(j)} \in \mathcal{P}_{ei}, n = |C_i|),$$

(14)

where $n$ is the sample number of cluster $C_i$.

The data of the $k$-th span in the cluster $C_i$ can be represented as:

$$\mathcal{P}_{C_i}^{(k)} = (x_{C_i}^{(k)}, \delta_{C_i}^{(k)}) \in C_i,$$

(15)

where $x_{C_i}^{(k)} = (p_i^{(k)}, e_i^{(k)}, b_i^{(k)}, r_i^{(k)}, s_i^{(k)}, a_i^{(k)}, f_i^{(k)}, h_i^{(k)}, l_i^{(k)}, t_i^{(k)}, \psi_i^{(k)})$, $\delta_{C_i}^{(k)} = \Delta f_{C_i}^{(k)}, k < |C_i|$.

3.4. Importance of the Features Used by the Model after Model Training

To reveal the relative importance of each feature when estimating and provide a better understanding of the data, the GBDT is introduced to estimate the relative importance of each feature. GBDT are tree ensemble algorithms. One of the advantages of tree ensemble algorithms is to output the importance of the features used by the model after model training. A tree can be formally expressed as:

$$T(x; \Theta) = \sum_{j=1}^{f} \gamma_j I(x \in R_j),$$

(16)

with parameters $\Theta = \{R_j, \gamma_j\}_{j=1}^f$, where $R_j, j = 1, 2, \ldots, f$ are the disjoint regions of the joint predictor variable space, $j$ is the terminal nodes of the tree; $\gamma_j$ is a constant assigned to each $R_j$; $I(S)$ is the indicator function of the set $S$ that maps elements of $S$ to 1 and all other elements to 0.
The boosted tree is a sum of trees,

\[ f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m), \tag{17} \]

where \( M \) is the number of trees.

For a single decision tree \( T \), the square of the relative importance of the variable \( X_l \) is the sum of the squared improvements of all internal nodes chosen as the splitting variable [48]:

\[ I^2_l(T) = \sum_{t=1}^{J-1} \hat{\tau}^2_t I(v(t) = l), \tag{18} \]

where the sum is over the \( J - 1 \) internal nodes of the tree, \( t \) is the node, the input variable \( X_{v(t)} \) divides the region into two subregions at node \( t \), \( \hat{\tau}^2_t \) is the maximal estimated improvement in squared error risk for the feature \( X_l \) over the entire region.

The global importance of feature \( X_l \) is measured by the average of the relative importance over trees:

\[ I^2_l = \frac{1}{M} \sum_{m=1}^{M} I^2_l(T_m). \tag{19} \]

In general, importance indicates the usefulness or value of each feature in building a boosted decision tree in the model. The more a feature uses a decision tree to make key decisions, the higher its relative importance. GBDT and XGBoost are tree ensemble algorithms, the feature importance can be calculated and sorted for each feature when the predictive model is trained.

### 3.5. HMCX Method for Sag Estimation

The advantage of the catenary model is that the resulting model is highly adaptable. XGBoost, being an ensemble learning technology, has a good gradient boosting performance in subtrees, and also exhibits good extrapolation capability for sparse data [32]. The wide-area sag estimation problem faced by the ALS operation scenario can be addressed more effectively with the two.

In this paper, we proposed HMCX to address the problem of sag estimation based on corridor data. The technical details are described below.

#### 3.5.1. The Catenary-Based Method

The catenary model (5) given in Section 3.2.1 is used to obtain a rough calculated sag value \( \hat{f}_i \) in the Equation (7).

According to the catenary model, the calculated sag of cluster \( C_i \) is:

\[ \hat{f}_{C_i} = \{ f^1_{C_i}, \ldots, f^n_{C_i} \} \times |C_i|). \tag{20} \]

The sag difference corresponding to the corridor dataset between the calculated sag and the sag value from LiDAR is:

\[ \delta_{C_i} = \{ \Delta f^1_{C_i}, \ldots, \Delta f^n_{C_i} \}, \Delta f^j_{C_i} \in \mathbb{R}, n = |C_i|). \tag{21} \]

#### 3.5.2. The Data-Driven Method

The ensemble learning algorithm XGBoost is used as the data-driven model in HMCX. The main principles of XGBoost are introduced in this section.

For the corridor dataset with \( n \) examples and 15 features \( P_{ei} = (x_i, \delta_i) \times |P_{ei}| = n < N, x_i \in \mathbb{R}^{14}, \delta_i \in \mathbb{R} \), the ensemble learning based on XGBoost are constructed [32].

\[ \delta_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}, \tag{22} \]
where $F = \{ f(x) = \omega_q(x) \} (q : \mathbb{R}^m \rightarrow \mathbb{T}, \omega \in \mathbb{R}^T)$ denotes the space of regression trees. Each $f_k$ represents an independent tree structure $q$ and leaf weights $\omega$.

Model complexity is introduced to measure the computational efficiency of the algorithm. The objective function of XGBoost is defined as:

$$
\mathcal{L}(\phi) = \sum_i l(\delta_i, \hat{\delta}_i) + \sum_k \Omega(f_k),
$$

$$
\Omega(f) = \gamma T + \frac{1}{2\lambda}||\omega||^2,
$$

where $l$ is a differentiable convex loss function that measures the difference between the prediction sag difference $\hat{\delta}_i$ and the target sag difference $\delta_i$, $\Omega$ represents the complexity of the model, and $k$ represents all the trees established.

The objective function is transformed into:

$$
\mathcal{L}^{(i)} = \sum_{i=1}^n l(\delta_i, \delta_i^{(t-1)} + f_i(x_i)) + \Omega(f_i)
$$

$$
\simeq \sum_{i=1}^n [g_if_i(x_i) + \frac{1}{2}h_if_i^2(x_i)] + \Omega(f_i),
$$

where $g_i = \partial_{\delta_i^{(t-1)}} l(e_i, \delta_i^{(t-1)})$ and $h_i = \partial^2_{\delta_i^{(t-1)}} l(e_i, \delta_i^{(t-1)})$ are the first and second derivatives on $l$. $\mathcal{L}^{(i)}$ presents the quality of a tree structure. The lower value of the objective function, the better the overall structure of the tree.

After a given sample split ratio, the dataset of each cluster is divided into training and test set. Then, the data-driven model is trained through utilizing the training set. The output and input of the training procedure are the test set information $x_{\mathcal{C}_i}$ and the sag differences $\delta_{\mathcal{C}_i}$ of the database $\mathcal{P}_c$. For cluster $\mathcal{C}_i$, performing the above operations for $\mathcal{P}_{\mathcal{C}_i} = (x_{\mathcal{C}_i}, \delta_{\mathcal{C}_i})$, then the trained data-driven model (22) (DDM #i) is created for cluster $\mathcal{C}_i$. Finally, the data-driven models for clusters based on sag differences is developed. After training, the data-driven models obtain the best parameters, and the trained model can be used to estimate the sag difference. The DDM #i predicts the sag difference $\hat{\delta}_{\mathcal{C}_i}$ of the data in cluster $\mathcal{C}_i$.

According to the data-driven model, the estimated sag difference of cluster $\mathcal{C}_i$ is:

$$
\hat{\delta}_{\mathcal{C}_i} = \{\hat{\delta}_{\mathcal{C}_i}, \ldots, \hat{\delta}_{\mathcal{C}_i}^n\} (n = |\mathcal{C}_i|).
$$

3.5.3. The HMCX Method

This section describes the HMCX procedure for sag estimation. For the $k$ span sag estimation in the cluster $\mathcal{C}_i$, the catenary model in (5) is used to determine the rough sag value $f_{\mathcal{C}_i}^k$ of the $k$-th span at first. Then, the sag difference $\delta_{\mathcal{C}_i}$ is predicted by performing the data-driven model of cluster $\mathcal{C}_i$. Finally, the sag estimation result is determined by adding the calculated sag $\hat{f}_{\mathcal{C}_i}$ and the sag difference $\delta_{\mathcal{C}_i}$ estimated by the data-driven model. The estimated sag $\hat{f}_{\mathcal{C}_i}$ of the $k$-th span in the cluster $\mathcal{C}_i$ is:

$$
\hat{f}_{\mathcal{C}_i}^k = f_{\mathcal{C}_i}^k + \delta_{\mathcal{C}_i}^k.
$$

Finally, for any span data $\mathcal{P}_l = (x_l, y_l)$, there is an estimated sag $\hat{f}_l$, and all datasets with estimated sag are:

$$
\mathcal{F} = \{\hat{f}_1, \ldots, \hat{f}_N\} (N = |\mathcal{P}|).
$$

The estimated sag for cluster $\mathcal{C}_i$ is:

$$
\hat{F}_{\mathcal{C}_i} = \{\hat{f}_{\mathcal{C}_i}^1, \ldots, \hat{f}_{\mathcal{C}_i}^n\} (n = |\mathcal{C}_i|).
$$
3.6. The Framework of HMCX

The HMCX framework for wide-area sag estimation is depicted in Figure 5, which combines the catenary-based model with the data-driven model.

![Diagram of the HMCX framework](image)

**Figure 5.** The HMCX framework for sag estimation.

The HMCX framework consists of two parts:

1. Offline training. The offline training phase mainly illustrates the procedure for HMCX training. First, the corridor database is split into several clusters utilizing the k-means method, and each cluster is divided into a test set and a training set. Next, the sag of the catenary line span is calculated utilizing test set information, and the sag difference to the real sag is determined from LiDAR. Then, the XGBoost model is trained by utilizing the sag differences and the test set information. Finally, the data-driven models for clusters based on sag differences are developed.

2. Hybrid model sag estimation. This phase describes the procedure for sag estimation implementing the HMCX. First, the test set data is clustered by utilizing the clustering model produced by offline training. Then, the catenary model is used to determine the
sag value of the data, and the sag difference prediction is performed based on the data-driven model to which it belongs. Finally, the sag estimation result is determined by adding the calculated sag and the sag difference estimated by the data-driven model.

**Remark 1.** The conversion between the HMCX and the XGBoost can be accomplished by controlling the catenary model’s output switch. When the catenary model’s output is 0, the HMCX is equivalent to the XGBoost model. This approach allows for more flexible model utilization in practice.

### 3.7. Performance Indicators

To prove the effectiveness and feasibility of the proposed methods, mean absolute error (MAE), root mean square error (RMSE), R-squared coefficient of determination ($R^2$), and Theil inequality coefficient ($TIC$) are selected as indicators.

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (f_i - \hat{f}_i)^2},
\]

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |f_i - \hat{f}_i|,
\]

\[
R^2 = 1 - \frac{\sum (f_i - \hat{f}_i)^2}{\sum (f_i - \bar{f})^2},
\]

\[
TIC = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (f_i - \hat{f}_i)^2}}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} f_i^2 + \frac{1}{N} \sum_{i=1}^{N} \hat{f}_i^2}},
\]

where $\hat{f}_i$ is the estimated value through utilizing HMCX, $f_i$ is the sag value extracted from LiDAR data, and $\bar{f}$ is the mean value of all sag $f_i$, where $\bar{f} = \frac{1}{N} \sum_{i=1}^{N} f_i$.

**Remark 2.** The RMSE is used to measure the deviation between the estimated sag and the sag value of the point cloud. The $R^2$ reflects the proportion of the dependent variable. The interval of $R^2$ is (0, 1). When $R^2$ is closer to 1, it means that the estimated value has a higher correlation with the actual value, that all estimations perfectly match the real results. TIC is between 0 and 1, the smaller the value, the smaller the difference, and the higher the estimation accuracy between the fitted value and the true value.

### 4. Experimental Results

In this section, the corridor database is used to verify the feasibility of the above method and processing framework. A total of 30,944 valid line data points were processed and obtained. The data processing and analysis platform were run on a Windows 10 computer with a 3.8 GHz Intel processor and 8 GB RAM.

#### 4.1. Results of Data Analysis

The correlation matrix shows the Pearson correlation coefficient between two variables, which is used to count the degree of linear correlation between two variables. The correlation matrix diagrams between features are shown in Figure 6.

According to the correlation matrix diagram, the influencing factors of the maximum sag of the large crossing were ranked according to the correlation coefficient: SPL (0.95), DW (0.11), TCSA (0.11), WPL (0.1), SCD (−0.094), SPT (−0.09), LEC (0.081), ELC (−0.075), TR (0.065), AT (0.055), BRF (0.054), WS (0.052), ST (−0.044), VTG (0.034), RPK (0.023), HD (0.0081). SPL is the most similar to MS and can achieve a correlation of 0.95. The HD, BRF, and ST have little correlation with maximum sag.
4.2. Results of Cluster Analysis

To reduce the influence of spatial sparsity and heterogeneity on the estimation results, we partition the corridor dataset with clustering methods to improve the similarity within a cluster. Six clustering algorithms (k-means, MeanShift, Ward, agglomerative clustering, DBSCAN, and Gaussian mixture) were introduced for comparative analysis. The performance of clustering methods was compared. Three common performance indicators, Calinski–Harabasz score (CHS), Silhouette score (SS), and Davies–Bouldin index (DBI), were selected to evaluate the effects of the above six clustering methods. The specific explanations of the indicators are omitted. The clustering performance results are shown in Table 3.

It can be seen from Table 3 that when the number of clusters was 3, the relative maximum value of SS was 0.735, and the relative minimum value of DBI was 0.374. Although the CHS index results obtained when the number of clusters was 10 appear to be better, its essence is caused by the high dispersion of the data. Therefore, a suboptimal solution was compromised, and the initial condition of k-means clustering was set to $k = 3$.

In order to compare the consistency of the results obtained by different clustering algorithms, this section introduces a Kappa-based consistency test to verify the consistency of the clustering results [49]. Table 4 shows the frequency of clusters under different clustering algorithms. Table 5 shows the Kappa consistency test results between each two algorithms.
Table 3. The performance indicators of six clustering methods with corridor data.

| Cluster Numbers | Indicators | k-Means | MeanShift | Ward | Agglomerative Clustering | DBSCAN | Gaussian Mixture |
|-----------------|------------|---------|-----------|------|-------------------------|--------|------------------|
| Clusters = 2    | CHS        | 78,061.734 | -       | 64,290.701 | 15,076.190 | - | 78,061.734 |
|                 | SS         | 0.712 | -       | 0.665 | 0.662 | - | 0.712 |
|                 | DBI        | 0.419 | -       | 0.468 | 0.414 | - | 0.419 |
|                 | Time       | 0.59 s | -       | 74.88 s | 70.73 s | - | 0.21 s |
| Clusters = 3    | CHS        | 100,452.646 | -       | 88,652.590 | 100,452.646 | - | 98,439.753 |
|                 | SS         | 0.735 | -       | 0.697 | 0.735 | - | 0.733 |
|                 | DBI        | 0.374 | -       | 0.480 | 0.374 | - | 0.374 |
|                 | Time       | 0.27 s | -       | 76.14 s | 60.73 s | - | 0.36 s |
| Clusters = 4    | CHS        | 102,776.263 | -       | 88,355.557 | 78,435.352 | - | 98,193.338 |
|                 | SS         | 0.692 | -       | 0.681 | 0.668 | - | 0.649 |
|                 | DBI        | 0.480 | -       | 0.464 | 0.412 | - | 0.531 |
|                 | Time       | 0.16 s | -       | 75.81 s | 65.40 s | - | 0.28 s |
| Clusters = 5    | CHS        | 128,894.975 | -       | 98,908.074 | 65,663.132 | - | 125,018.748 |
|                 | SS         | 0.669 | -       | 0.649 | 0.666 | - | 0.658 |
|                 | DBI        | 0.506 | -       | 0.489 | 0.311 | - | 0.482 |
|                 | Time       | 0.19 s | -       | 77.91 s | 66.60 s | - | 0.35 s |
| Clusters = 10   | CHS        | 360,813.746 | 305,029.967 | 496,817.066 | 305,029.967 | - | 440,830.328 |
|                 | SS         | 0.802 | 0.805 | 0.831 | 0.805 | - | 0.839 |
|                 | DBI        | 0.287 | 0.237 | 0.343 | 0.237 | - | 0.309 |
|                 | Time       | 0.35 s | 0.42 s | 80.71 s | 65.63 s | - | 0.70 s |
| Clusters = 78   | CHS        | -       | -       | -       | 3.46978973 | - | - |
|                 | SS         | -       | -       | -       | -0.88909178 | - | - |
|                 | DBI        | -       | -       | -       | 1.742356117 | - | - |
|                 | Time       | -       | -       | -       | 10.56 s | - | - |

1 The optimal number of clusters for the Meanshift algorithm is 10, and indicators under other cluster numbers are not considered. 2 This is the result after DBSCAN clustering, other algorithms do not consider the indicators when the clusters are 78.

Table 4. Frequency of clusters under different clustering algorithms.

| Method 1 | Cluster | Frequency | Percentage (%) | Cumulative Percentage (%) |
|----------|---------|-----------|----------------|----------------------------|
| k-means  | 0       | 11,104    | 35.88          | 35.88                      |
|          | 1       | 16,981    | 54.88          | 90.76                      |
|          | 2       | 2859      | 9.24           | 100                        |
| Ward     | 0       | 11,789    | 38.1           | 38.1                       |
|          | 1       | 16,396    | 52.99          | 91.08                      |
|          | 2       | 2759      | 8.92           | 100                        |
| Gaussian Mixture | 0       | 11,958    | 38.64          | 38.64                      |
|          | 1       | 17,277    | 55.83          | 94.48                      |
| MeanShift | 0       | 30,572    | 98.8           | 98.8                       |
|          | 1       | 372       | 1.2            | 100                        |
| Agglomerative Clustering | 0       | 30,387    | 98.2           | 98.2                       |
|          | 1       | 405       | 1.31           | 99.51                      |
|          | 2       | 152       | 0.49           | 100                        |
| Total    | 30,944  | 100       | 100            | 100                        |

1 The DBSCAN algorithm was skipped because no valid clusters could be obtained.
Table 5. Kappa consistency test results of clustering results.

| Item                                      | Kappa Value | z-Value | p-Value | Standard Error | 95% CI \(^2\) |
|-------------------------------------------|-------------|---------|---------|----------------|----------------|
| k-means & Ward                            | 0.868       | 188.304 | 0.000 **| 0.003          | 0.863–0.873    |
| k-means & Gaussian Mixture                | 0.881       | 186.135 | 0.000 **| 0.002          | 0.876–0.886    |
| k-means & MeanShift                       | 0.015       | 16.838  | 0.000 **| 0.001          | 0.014–0.017    |
| k-means & Agglomerative Clustering        | 0.0026      | 25.892  | 0.000 **| 0.001          | 0.024–0.029    |
| Ward & Gaussian Mixture                   | 0.889       | 187.034 | 0.000 **| 0.002          | 0.884–0.894    |
| Ward & MeanShift                          | 0.017       | 17.519  | 0.000 **| 0.001          | 0.015–0.018    |
| Ward & Agglomerative Clustering           | 0.028       | 26.664  | 0.000 **| 0.001          | 0.026–0.031    |
| Gaussian Mixture & MeanShift              | 0.016       | 16.797  | 0.000 **| 0.001          | 0.015–0.018    |
| Gaussian Mixture & Agglomerative Clustering | 0.028     | 26.42   | 0.000 **| 0.001          | 0.026–0.031    |
| MeanShift & Agglomerative Clustering      | 0.439       | 88.471  | 0.000 **| 0.02           | 0.399–0.479    |

1 **p < 0.01. 2 It means the 95% CI value of Fleiss Kappa. Its value = Fleiss Kappa value +/– 1.9599638454005 × standard error/n, where n is the number of rows.

From the cluster frequency in Table 4 and the Kappa value in Table 5, it can be seen that k-means, Ward, and Gaussian Mixture show strong consistency in the clustering effect, and the clustering frequency is roughly similar. From a statistical point of view, the \( p < 0.01 \), which means there is a significant correlation between the clustering results. MeanShift and Agglomerative Clustering are less consistent with other algorithms, these two algorithms perform poorly on the corridor dataset and fail to segment the dataset well.

After comprehensively considering the performance of each index and the computational complexity, the following results were obtained: k-means had higher SS and lower DBI and outperformed other methods on corridor data. The poor performance of the density-based DBSCAN clustering method on the inspection dataset also proves the sparseness of the data. The best performance was obtained when the corridor data was divided into three clusters. Therefore, we chose k-means as the clustering method, and set the number of clusters to three. Figure 7 shows the distribution of clustering results on the span length. There are three clusters and their frequencies were 11,610, 18,228, and 1106, respectively.

![Figure 7. Span length distribution of 3 clusters.](image)

After the k-means clustering was completed, the feature importance to the sag difference in the three clusters was calculated and ranked based on GBDT. The proportion of feature importance to the sag difference in the clusters is shown in Figure 8.

The first three most important variables for the sag difference value from Cluster #0 in the Figure 8 are DW, WPL, SCD. For Cluster #1, the first three most important variables are WPL, DW, LEC. The first three most important variables for Cluster #2 are ELF, BRF, ST.
According to the actual operation and maintenance experience, the lines corresponding to these three clusters may be: large-span lines with larger diameter conductors, heavy-duty lines with larger ampacities, and aging lines with long-term service.

![Figure 8](image)

**Figure 8.** The proportion of feature importance to the sag difference in the clusters.

### 4.3. Estimation Result Analysis

To verify the effectiveness of HMCX, the catenary model and three data-driven models based on XGBoost, Linear Regression (LR), and Bayesian Ridge Regression (BayesRR) were introduced for comparative analysis. At the same time, to verify the clustering effect, the indicators of the above methods for all data and each cluster are presented. In order to obtain a reliable and stable model and obtain data results relatively objectively, 10-fold cross-validation was used to partition the dataset. The dataset was randomly divided into 10 parts, of which 9 parts were used for model training and the remaining 1 part was used for testing. We repeated this process 10 times, keeping different remainders for testing.

For all data, the 100 example results of the test set data were randomly extracted for display. The estimation results and estimation error of the data based on the above methods are shown in Figures 9 and 10. The frequency distributions of the estimation error of HMCX are shown in Figure 11.

From Figures 10 and 11, it can be seen that the overall error fluctuation and error distribution of HMCX are smaller than other methods for all data. The catenary has the largest positive error fluctuation, up to 20 m, which may be caused by the high uncertainty of the field measurement of its related parameters. Although HMCX achieves smaller errors than other methods, it can be clearly seen that there is a mutation data at the position where the span number is 20.
Figure 9. Sag estimation results of all data.

Figure 10. Sag estimation errors of all data.

Figure 11. The frequency distributions of sag error for all data.
In order to see the effect of clustering more intuitively, we compared the three clusters, and the following are their results. The sag estimation results and the sag estimation errors of cluster #0 are shown in Figures 12 and 13. The sag estimation results and the errors of cluster #1 are shown in Figures 14 and 15. And the results of cluster #2 are shown in Figures 16 and 17.

From Figures 12 and 13, it can be seen that the effect of HMCX on cluster #0 is significantly better than other methods. At the same time, it can be seen that both the catenary and LR have large error fluctuations.

From Figures 14 and 15, it can be seen that the error performance of LR on cluster #1 is worse than that of the catenary. The error fluctuations of HMCX, BayesRR, and XGBoost on cluster #1 are smaller.

![Figure 12. Sag estimation results of cluster #0.](image1)

![Figure 13. Sag estimation errors of cluster #0.](image2)
Figure 14. Sag estimation results of cluster #1.

Figure 15. Sag estimation Errors of cluster #1.

Figure 16. Sag estimation results of cluster #2.
It can be seen from Figures 16 and 17 that the error of the catenary on cluster #2 presents a reverse positive error, and the error of the catenary model is close to 20 m. The catenary model using the fixed horizontal stress is not suitable for the cluster. The HMCX#2 can well compensate for the errors caused by the catenary, and effectively improve the accuracy of the sag estimation of the cluster. The effects of HMCX, BayesRR, and XGBoost on cluster #2 are also acceptable.

For all data and three different clusters, the evaluation indicators of the above methods are shown in Table 6.

| Category | Models     | RMSE   | MAE    | TIC   | $R^2$  | Time   |
|----------|------------|--------|--------|-------|--------|--------|
| All data | Catenary   | 2.667  | 1.315  | 0.072 | 0.952  | 8.223 s|
|          | HMCX       | 0.715  | 0.433  | 0.019 | 0.996  | 12.752 s|
|          | XGBoost    | 0.931  | 0.402  | 0.025 | 0.994  | 8.336 s |
|          | LR         | 3.384  | 2.571  | 0.092 | 0.923  | 0.119 s |
|          | BayesRR    | 3.383  | 2.372  | 0.091 | 0.924  | 0.048 s |

Cluster #0

| Models     | RMSE   | MAE    | TIC   | $R^2$  | Time   |
|------------|--------|--------|-------|--------|--------|
| Catenary   | 2.902  | 1.388  | 0.068 | 0.955  | 3.226 s|
| HMCX#0     | 0.484  | 0.322  | 0.011 | 0.999  | 5.309 s |
| XGBoost    | 0.443  | 0.273  | 0.010 | 0.999  | 3.123 s |
| LR         | 3.575  | 2.457  | 0.085 | 0.932  | 0.075 s |
| BayesRR    | 3.571  | 2.461  | 0.085 | 0.933  | 0.034 s |

Cluster #1

| Models     | RMSE   | MAE    | TIC   | $R^2$  | Time   |
|------------|--------|--------|-------|--------|--------|
| Catenary   | 1.799  | 0.950  | 0.052 | 0.976  | 4.501 s|
| HMCX#1     | 0.581  | 0.383  | 0.017 | 0.997  | 7.622 s |
| XGBoost    | 0.775  | 0.370  | 0.022 | 0.995  | 4.022 s |
| LR         | 3.182  | 2.341  | 0.092 | 0.924  | 0.009 s |
| BayesRR    | 3.183  | 2.342  | 0.092 | 0.924  | 0.018 s |

Cluster #2

| Models     | RMSE   | MAE    | TIC   | $R^2$  | Time   |
|------------|--------|--------|-------|--------|--------|
| Catenary   | 9.285  | 7.172  | 0.491 | −0.156 | 0.258 s|
| HMCX#2     | 1.345  | 0.864  | 0.046 | 0.976  | 0.561 s |
| XGBoost    | 0.792  | 0.423  | 0.028 | 0.992  | 0.421 s |
| LR         | 3.072  | 1.866  | 0.111 | 0.874  | 0.004 s |
| BayesRR    | 3.085  | 1.862  | 0.112 | 0.872  | 0.006 s |

From Table 6, the weak performance of the catenary can be seen from the $R^2$ of $-0.156$, which is also illustrated by the same TIC of 0.491. The catenary model may not be suitable for all lines due to parameter drift and variable uncertainty. HMCX can significantly improve the estimation performance compared to the catenary. For all data, HMCX has the smallest RMSE. In both clusters #0 and #1, HMCX achieves better performance than the full data. In cluster #2, the RMSE of HMCX was increased due to the large error of the...
catenary model, but the estimation performance was still better than that of LR, ByesRR, and catenary. In terms of time performance, since HMCX includes XGBoost and catenary, its running time is longer, but still acceptable. In terms of overall estimation performance, HMCX significantly outperforms catenary, LR, and BayesRR.

5. Conclusions

This paper uses corridor data to solve the problem of wide-area sag estimation in the power inspection field. A systematic data processing and analysis framework for aircraft-based inspection is constructed to preprocess multi-source data. The proposed HMCX method combines the adaptability of catenary with the sparsity awareness of XGBoost. The feasibility and effectiveness of the proposed HMCX are verified by practical data. The proposed HMCX method outperforms catenary, LR, and BayesRR involved in this study and shows a promising prospect in wide-area sag estimation of the power dispatching and inspection.

According to the above analysis results, it is effective to reduce the heterogeneity of corridor data by improving the similarity between data by clustering. The accuracy of wide-area sag estimation is improved compared to the catenary. However, from the perspective of performance indicators, the impact of the catenary on the HMCX cannot be ignored. Therefore, the future work can be considered from the following aspects:

1. The reasons for the errors of the catenary model in different clusters need to be found and eliminated through further analysis of the line clustering.
2. More suitable model parameters or models can be selected according to different clustering features, to reduce the influence of the calculation bias of the mechanism part on the estimation.
3. The recommendation algorithm can be used to impute the missing data of wide-area lines, improve the matching accuracy of span similarity, and reduce the impact of data heterogeneity.
4. Heuristic optimization algorithms can be introduced to transform subset selection into an optimization problem to find the optimal subset of features for the model.

Author Contributions: Conceptualization, Y.W. and B.Z.; methodology, Y.W.; validation, Y.-H.L., Y.W. and B.Z.; formal analysis, C.-Y.S.; investigation, A.M.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, Y.W., C.-Y.S. and B.Z.; visualization, Y.W. and Y.-H.L.; supervision, C.-Y.S.; project administration, B.Z.; funding acquisition, B.Z. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Project of National Natural Science Foundation of China grant number 62173099, the Guangdong Power Grid Co., Ltd. Science and Technology Project under Grant 031000KK52180007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Beyond the authors of the paper, I would like to express my special thanks to my senior Jun-yi Li, my schoolmate Wenshuai Lin, and my junior Zhengganzhe Chen. They provided invaluable advice on my research and helped me during difficult times. Their help greatly contributed to the smooth completion of the project. Besides, many thanks of gratitude to my industrial advisor Huamin Zhou, and all the leaders who helped me. They gave me the opportunity to participate in grid technology projects and provided valuable information. I would also like to thank my family and friends for their support. I really appreciate them.

Conflicts of Interest: The authors declare no conflict of interest.
Abbreviations

The following abbreviations are used in this manuscript:

- ALS: Airborne Laser Scanning
- LiDAR: Light Detection And Ranging
- PMU: Phasor Measurement Unit
- HMCX: Hybrid Model based on Catenary and XGBoost
- VTG: Voltage Level
- SPT: Span Type
- CDT: Conductor Type
- TR: Terrain
- ST: Service Time
- SPL: Span Length
- HD: Height Difference
- DMT: Distance from the Maximum Sag Point to the Tower
- MS: Maximum Sag
- AT: Ambient Temperature
- WS: Wind Speed
- ELC: Elastic Coefficient
- BRF: Breaking Force
- RPK: Resistance Per kilometer
- DW: Diameter of Wire
- LEC: Linear Expansion Coefficient
- WPL: Weight Per unit Length
- TCSA: Total Cross Sectional Area
- SCD: Steel Core Diameter
- GBDT: Gradient Boosting Decision Tree
- PCA: Principal Component Analysis
- XGBoost: eXtreme Gradient Boosting
- DDM: Data-Driven Model
- MAE: Mean Absolute Error
- RMSE: Root Mean Square Error
- $R^2$: R-Squared Coefficient of Determination
- TIC: Theil Inequality Coefficient
- ETL: Extract, Transform, and Load
- DBSCAN: Density-Based Spatial Clustering of Applications with Noise
- CHS: Calinski–Harabasz Score
- SS: Silhouette Score
- DBI: Davies–Bouldin Index
- LR: Linear Regression
- BayesRR: Bayesian Ridge Regression
- CPF: Cubic Polynomial Fitting

References

1. Chen, Y.; Lin, J.; Liao, X. Early detection of tree encroachment in high voltage powerline corridor using growth model and UAV-borne LiDAR. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102740. [CrossRef]
2. Ortega, S.; Trujillo, A.; Santana, J.M.; Suárez, J.P.; Santana, J. Characterization and modeling of power line corridor elements from LiDAR point clouds. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 24–33. [CrossRef]
3. Yue, C.D.; Chiu, Y.S.; Tu, C.C.; Lin, T.H. Evaluation of an offshore wind farm by using data from the weather station, floating LiDAR, mast, and MERRA. *Energies* **2020**, *13*, 185. [CrossRef]
4. Meng, A.; Li, J.; Yin, H. An efficient crisscross optimization solution to large-scale non-convex economic load dispatch with multiple fuel types and valve-point effects. *Energy* **2016**, *113*, 1147–1161. [CrossRef]
5. Douglass, D.A.; Gentle, J.; Nguyen, H.M.; Chisholm, W.; Xu, C.; Goodwin, T.; Chen, H.; Nuthalapati, S.; Hurst, N.; Grant, J.; et al. A review of dynamic thermal line rating methods with forecasting. *IEEE Trans. Power Deliv.* **2019**, *34*, 2100–2109. [CrossRef]
6. Safdarian, A.; Degefa, M.Z.; Fotuhi-Firuzabad, M.; Lehtonen, M. Benefits of real-time monitoring to distribution systems: Dynamic thermal rating. *IEEE Trans. Smart Grid* **2015**, *6*, 2023–2031. [CrossRef]
7. Polevoy, A. Impact of data errors on sag calculation accuracy for overhead transmission line. *IEEE Trans. Power Deliv.* **2014**, *29*, 2040–2045. [CrossRef]
8. Fan, F.; Bell, K.; Infield, D. Transient-state real-time thermal rating forecasting for overhead lines by an enhanced analytical method. *Electr. Power Syst. Res.* **2019**, *167*, 213–221. [CrossRef]

9. Sun, X.; Huang, Q.; Hou, Y.; Jiang, L.; Pong, P.W. Noncontact operation-state monitoring technology based on magnetic-field sensing for overhead high-voltage transmission lines. *IEEE Trans. Power Deliv.* **2013**, *28*, 2145–2153. [CrossRef]

10. Hajeforosh, S.; Bollen, M.H. Uncertainty analysis of stochastic dynamic line rating. *Electr. Power Syst. Res.* **2021**, *194*, 107043. [CrossRef]

11. Esfahani, M.M.; Yousefi, G.R. Real-time congestion management in power systems considering quasi-dynamic thermal rating and congestion clearing time. *IEEE Trans. Ind. Inform.* **2016**, *12*, 745–754. [CrossRef]

12. Mahajan, S.M.; Singareddy, U.M. A real-time conductor sag measurement system using a differential GPS. *IEEE Trans. Power Deliv.* **2012**, *27*, 475–480. [CrossRef]

13. Pan, L.; Xiao, X. Image recognition for on-line vibration monitoring system of transmission line. In Proceedings of the 2009 9th International Conference on Electronic Measurement & Instruments, Beijing, China, 16–19 August 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 3–1078.

14. Albizu, I.; Fernandez, E.; Eguia, P.; Torres, E.; Mazon, A.J. Tension and ampacity monitoring system for overhead lines. *IEEE Trans. Power Deliv.* **2012**, *28*, 3–10. [CrossRef]

15. Alvarez, D.L.; da Silva, F.F.; Mombello, E.E.; Bak, C.L.; Rosero, J.A. Conductor temperature estimation and prediction at thermal transient state in dynamic line rating application. *IEEE Trans. Power Deliv.* **2018**, *33*, 2236–2245. [CrossRef]

16. Wydra, M.; Ksala, P.; Harasim, D.; Kacejko, P. Overhead transmission line sag estimation using a simple optomechanical system with chirped fiber bragg gratings. Part 1: Preliminary measurements. *Sensors* **2018**, *18*, 309. [CrossRef]

17. Xu, Q.; Liu, X.; Zhu, K.; Pong, P.W.; Liu, C. Magnetic-field-sensing-based approach for current reconstruction, sag detection, and inclination detection for overhead transmission system. *IEEE Trans. Magn.* **2019**, *55*, 4003307. [CrossRef]

18. Kopsidas, K.; Rowland, S.M.; Boumeid, B. A holistic method for conductor ampacity and sag computation on an OHL structure. *IEEE Trans. Power Deliv.* **2012**, *27*, 1047–1054. [CrossRef]

19. Du, Y.; Liao, Y. On-line estimation of transmission line parameters, temperature and sag using PMU measurements. *Electr. Power Syst. Res.* **2012**, *93*, 39–45. [CrossRef]

20. Chen, C.; Yang, B.; Song, S.; Peng, X.; Huang, R. Automatic clearance anomaly detection for transmission line corridors utilizing UAV-Borne LIDAR data. *Remote Sens.* **2018**, *10*, 613. [CrossRef]

21. Golinelli, E.; Perini, U.; Barberis, F.; Musazzi, S. A Laser Scanning System for Sag Detection on the Overhead Power Lines: In Field Measurements. In *Sensors*; Springer: New York, NY, USA, 2014; pp. 311–314.

22. Le Clainche, S.; Lorente, L.S.; Vega, J.M. Wind predictions upstream wind turbines from a LiDAR database. *Energies* **2018**, *11*, 543. [CrossRef]

23. Guan, H.; Sun, X.; Su, Y.; Hu, T.; Wang, H.; Wang, H.; Peng, C.; Guo, Q. UAV-lidar aids automatic intelligent powerline inspection. *Int. J. Electr. Power Energy Syst.* **2021**, *130*, 106987. [CrossRef]

24. Awrangjeb, M. Extraction of power line pylons and wires using airborne lidar data at different height levels. *Remote Sens.* **2019**, *11*, 1798. [CrossRef]

25. Palmer, D.; Koumpli, E.; Cole, I.; Gottschalg, R.; Betts, T. A GIS-based method for identification of wide area rooftop suitability for minimum size PV systems using LiDAR data and photogrammetry. *Energies* **2018**, *11*, 3506. [CrossRef]

26. Shang, X.; Li, Z.; Zheng, J.; Wu, Q. Equivalent modeling of active distribution network considering the spatial uncertainty of renewable energy resources. *Int. J. Electr. Power Energy Syst.* **2019**, *112*, 83–91. [CrossRef]

27. Du, P.; Bai, X.; Tan, K.; Xue, Z.; Samat, A.; Xia, J.; Li, E.; Su, H.; Liu, W. Advances of four machine learning methods for spatial data handling: A review. *J. Geovisualization Spat. Anal.* **2020**, *4*, 1–25. [CrossRef]

28. Tien Bui, D.; Tuan, T.A.; Klempe, H.; Pradhan, B.; Revhaug, I. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* **2016**, *13*, 361–378. [CrossRef]

29. Schratt, P.; Muenchow, J.; Iturritxa, E.; Richter, J.; Brenning, A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol. Model.* **2019**, *406*, 109–120. [CrossRef]

30. Fouedjio, F.; Klump, J. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environ. Earth Sci.* **2019**, *78*, 1–24. [CrossRef]

31. Stefenon, S.F.; Ribeiro, M.H.D.M.; Nied, A.; Yow, K.C.; Mariani, V.C.; dos Santos Coelho, L.; Seman, L.O. Time series forecasting using ensemble learning methods for emergency prevention in hydroelectric power plants with dam. *Electr. Power Syst. Res.* **2022**, *202*, 107584. [CrossRef]

32. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

33. Khan, W.; Walker, S.; Zeiler, W. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy* **2022**, *240*, 122812. [CrossRef]

34. Xiao, F.; Ai, Q. Data-driven multi-hidden markov model-based power quality disturbance prediction that incorporates weather conditions. *IEEE Trans. Power Syst.* **2018**, *34*, 402–412. [CrossRef]

35. Xue, P.; Jiang, Y.; Zhou, Z.; Chen, X.; Fang, X.; Liu, J. Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms. *Energy* **2019**, *188*, 116085. [CrossRef]
36. Ågren, A.M.; Larson, J.; Paul, S.S.; Laudon, H.; Lidberg, W. Use of multiple LIDAR-derived digital terrain indices and machine learning for high-resolution national-scale soil moisture mapping of the Swedish forest landscape. *Geoderma* **2021**, *404*, 115280. [CrossRef]

37. Marrs, J.; Ni-Meister, W. Machine learning techniques for tree species classification using co-registered LiDAR and hyperspectral data. *Remote Sens.* **2019**, *11*, 819. [CrossRef]

38. Neuville, R.; Bates, J.S.; Jonard, F. Estimating forest structure from UAV-mounted LiDAR point cloud using machine learning. *Remote Sens.* **2021**, *13*, 352. [CrossRef]

39. Tan, J.; Zhao, H.; Yang, R.; Liu, H.; Li, S.; Liu, J. An Entropy-Weighting Method for Efficient Power-Line Feature Evaluation and Extraction from LiDAR Point Clouds. *Remote Sens.* **2021**, *13*, 3446. [CrossRef]

40. Maxwell, A.E.; Sharma, M.; Kite, J.S.; Donaldson, K.A.; Thompson, J.A.; Bell, M.L.; Maynard, S.M. Slope failure prediction using random forest machine learning and lidar in an eroded folded mountain belt. *Remote Sens.* **2020**, *12*, 486. [CrossRef]

41. Lv, B.; Xu, H.; Wu, J.; Tian, Y.; Zhang, Y.; Zheng, Y.; Yuan, C.; Tian, S. LiDAR-enhanced connected infrastructures sensing and broadcasting high-resolution traffic information serving smart cities. *IEEE Access* **2019**, *7*, 79895–79907. [CrossRef]

42. Li, W.; Luo, Z.; Xiao, Z.; Chen, Y.; Wang, C.; Li, J. A GCN-based method for extracting power lines and pylons from airborne LiDAR data. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5700614. [CrossRef]

43. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–10.

44. Graham, B.; Engelcke, M.; Van Der Maaten, L. 3d semantic segmentation with submanifold sparse convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9224–9232.

45. Li, N.; Kahler, O.; Pfeifer, N. A comparison of deep learning methods for airborne lidar point clouds classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6467–6486. [CrossRef]

46. Schmohl, S.; Sörgel, U. Submanifold sparse convolutional networks for semantic segmentation of large-scale ALS point clouds. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *4*, 77–84. [CrossRef]

47. Winiwarter, L.; Mandlburger, G.; Schmohl, S.; Pfeifer, N. Classification of ALS point clouds using end-to-end deep learning. *PFG—J. Photogramm. Remote Sens. Geoinf. Sci.* **2019**, *87*, 75–90. [CrossRef]

48. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 2017.

49. Fleiss, J.L.; Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619. [CrossRef]