


## Article

# Enhancing Zero-Carbon Building Operation and Maintenance: A Correlation-Based Data Mining Approach for Database Analysis

Yuhong Zhao <sup>1,2</sup>, Ruirui Liu <sup>1,2</sup>, Zhansheng Liu <sup>1,2,\*</sup> , Yun Lu <sup>3</sup>, Liang Liu <sup>1,2</sup>, Jingjing Wang <sup>1,2</sup> and Wenxiang Liu <sup>1,2</sup>

<sup>1</sup> Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, Beijing 100124, China

<sup>2</sup> Key Laboratory of Urban Security and Disaster Engineering of Ministry of Education, Beijing University of Technology, Beijing 100124, China

<sup>3</sup> Shangxinzhuan Canal Management Office in Huangzhong District, Xining 811600, China

\* Correspondence: liuzhansheng@bjut.edu.cn

**Abstract:** In the context of global climate change and the increasing focus on carbon emissions, carbon emission research has become a prominent area of study. However, research in this field inevitably involves extensive monitoring, and when the data become complex and chaotic, the accuracy of these data can be challenging to control, making it difficult to determine their reliability. This article starts by exploring the operational and maintenance data of zero-carbon buildings, aiming to uncover the correlation between energy consumption data and environmental data. This correlation is categorized into two main types: linear correlation and trend correlation. By establishing error degree calculations based on these correlation relationships, anomaly detection can be performed on the data. Analyzing the interrelationships between these datasets allows for the formulation of appropriate fitting equations, primarily consisting of linear and polynomial fits, all of which exhibit a determination coefficient exceeding 0.99. These fitting equations are then utilized to correct errors in the anomalous data, and the reasonableness of the fitting methods is demonstrated by examining the residual distribution. The final results align with the corresponding expectations, providing a concise and effective correction method for monitoring data in zero-carbon smart buildings. Importantly, this method exhibits a certain level of generality and can be applied to various scenarios within the realm of zero-carbon buildings.

**Keywords:** self-regulation; zero-carbon building; data mining; correlation analysis; fitting function



**Citation:** Zhao, Y.; Liu, R.; Liu, Z.; Lu, Y.; Liu, L.; Wang, J.; Liu, W.

Enhancing Zero-Carbon Building Operation and Maintenance: A Correlation-Based Data Mining Approach for Database Analysis. *Sustainability* **2023**, *15*, 13671.

<https://doi.org/10.3390/su151813671>

Academic Editor: Wen-Hsien Tsai

Received: 15 August 2023

Revised: 2 September 2023

Accepted: 9 September 2023

Published: 13 September 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the rapid acceleration of global industrialization and urbanization, the release of greenhouse gases into the atmosphere has intensified, leading to severe environmental issues [1]. As of the end of 2022, the average concentration of carbon dioxide in the air reached a concerning 417 ppm [2], necessitating immediate action to address these environmental challenges. Notably, construction activities alone contribute to 36% of worldwide energy consumption and a significant 39% of global carbon dioxide emissions [2]. In this context, the emergence and advancement of zero-carbon buildings offer a pivotal reference point for mitigating carbon emissions within the construction industry.

Currently, a substantial number of buildings are equipped with building automation control systems, with energy management being a key consideration. Gathering energy consumption-related data plays a vital role in realizing intelligent energy consumption control. However, the collected energy consumption data may not always be flawless [3]. On one hand, issues with the measuring instruments themselves, such as aging and delays, can lead to inaccuracies in the data. On the other hand, achieving high precision in

measurements often demands higher costs, making it impractical to require all measuring instruments to meet stringent standards of accuracy. In this context, it becomes essential to carefully examine and mine the source data to extract valuable insights and make informed decisions.

Some researchers have achieved significant improvements in the accuracy of building operational data by employing optimized prediction algorithms. Fu et al. [3] proposed a methodology based on a correlation coefficient and a wavelet-based support vector machine (SVM) predictor to detect and recover the proportional deviation data faults and faults caused by network communication. Ma et al. [4] propose a synchronous prediction method for predicting building energy consumption in the secondary branch, and, in this model, a synchronous data feature similarity (SDFS) model is used to find a similar energy consumption feature and an extreme gradient boosting (XGBoost) model is used for training and the production of accurate prediction results. Lukas Lundström [5] proposed an adaptive weather correction framework for energy consumption data, which used weather data as the input of the training model to predict energy consumption. The proposed method is more accurate for low-energy and net-zero-energy buildings. Alghamdi and Javaid [6] significantly enhanced the accuracy of subsequent predictions by preprocessing data related to smart grid loads and prices.

Another group of researchers has achieved favorable results in data preprocessing by combining theoretical approaches with complementary application algorithms. Chen et al. [7] present a sorted Top-N AD mechanism to generate a list of suspicious anomalous SMs and an error estimation model (EEM) using only SM electricity consumption data is investigated. The truncated singular value decomposition regularization with L-curve optimization (TSVD+L) method is proposed to address the model's ill-posedness. Choi et al. [8] have demonstrated the advantages of preprocessing data using a density-based circular temporal clustering method. This approach effectively identifies anomalies in year-long building energy data even without the assistance of domain-specific knowledge. Zach et al. [9] have outlined a scalable approach to building monitoring and data processing that is independent of suppliers and technologies. This distributed software architecture, achieved through robust data preprocessing algorithms, virtual data points, automatic building model calibration [10], and various software interfaces, allows for scalable handling of data streams required for a wide range of applications from individual buildings to city-level contexts. Bhagat et al. [11] introduced an exclusive framework for data preprocessing and data wrangling in which most of the helper functions are heavily used in every data preprocessing exercise and "Sparx" comes as a complete kit, irrespective of structured and unstructured data.

Based on the current research, studies on energy consumption measurement errors have predominantly focused on the data themselves, often overlooking the potential influence of surrounding environmental factors. Concurrently, there have been related research efforts aiming to improve energy consumption prediction accuracy by incorporating weather conditions and employing intelligent prediction algorithms for training. Although these endeavors have shown some improvements in prediction results, addressing data errors at the correction level remains a challenge. On the one hand, accurate prediction demands a substantial number of data, and obtaining all the corresponding data is sometimes unfeasible. On the other hand, detecting and diagnosing abnormal data is a critical aspect, requiring a rational and effective data diagnosis method. As a follow-up study, the center of focus is placed on the measurement data related to energy consumption in zero-carbon buildings, with a particular emphasis on exploring potential correlations among energy consumption-related data. To address the issue of measurement errors in zero-carbon data, a concise yet effective method for detecting anomalies in energy consumption data has been established. By thoroughly analyzing the underlying relationships among various factors, the aim is to identify the most suitable fitting function for updating and rectifying abnormal values.

## 2. Enhancing Zero-Carbon Building Operation and Maintenance: Database Development, Data Mining Process, and Targeted Services

Aiming at the energy consumption operation characteristics of zero-carbon buildings, this research constructs a general database model of zero-carbon buildings with reference to relevant papers, and presents the operation mode of the data display platform based on digital twin technology [12]. The digital twin platforms' data processing flow encompasses eight modules: data collection, transmission, storage, preprocessing, processing, retrieval, feedback, and control (Figure 1). Data collection is typically accomplished using BIM models and Internet of Things (IoT) systems to gather data that are both currently utilized and potentially useful in the day-to-day operation and maintenance of zero-carbon buildings. This stage serves as a fundamental underpinning for the smooth operation of the entire maintenance and management platform. The collected data are then transmitted in real time to local servers and cloud servers via transmission systems such as intelligent gateways. These initial datasets often present issues like data loss and anomalies and are generally not directly processed or utilized. Instead, they require preprocessing, which involves further screening and integration [13] to extract pertinent information. Subsequently, certain data points, such as real-time power generation, electricity consumption, and water consumption, are directly accessed by the operation and maintenance platform to provide users with relevant information. Additionally, some information, including historical trends in electricity and water consumption, personnel flow patterns, and even abnormal operational statuses throughout the entire building, is obtained through the analysis and processing of the corresponding datasets. This information is indirectly accessed by the operation and maintenance platform to serve as a basis for informed decision making regarding the overall building operation strategy. The process of indirect data retrieval is more intricate, but both direct and indirect data utilization rely on the critical data provided during the preprocessing step. Finally, the information retrieved through both direct and indirect means is fed back to the intelligent control terminal. This feedback loop enables the control of the fundamental operation and maintenance aspects of zero-carbon buildings, thus establishing a closed-loop system.

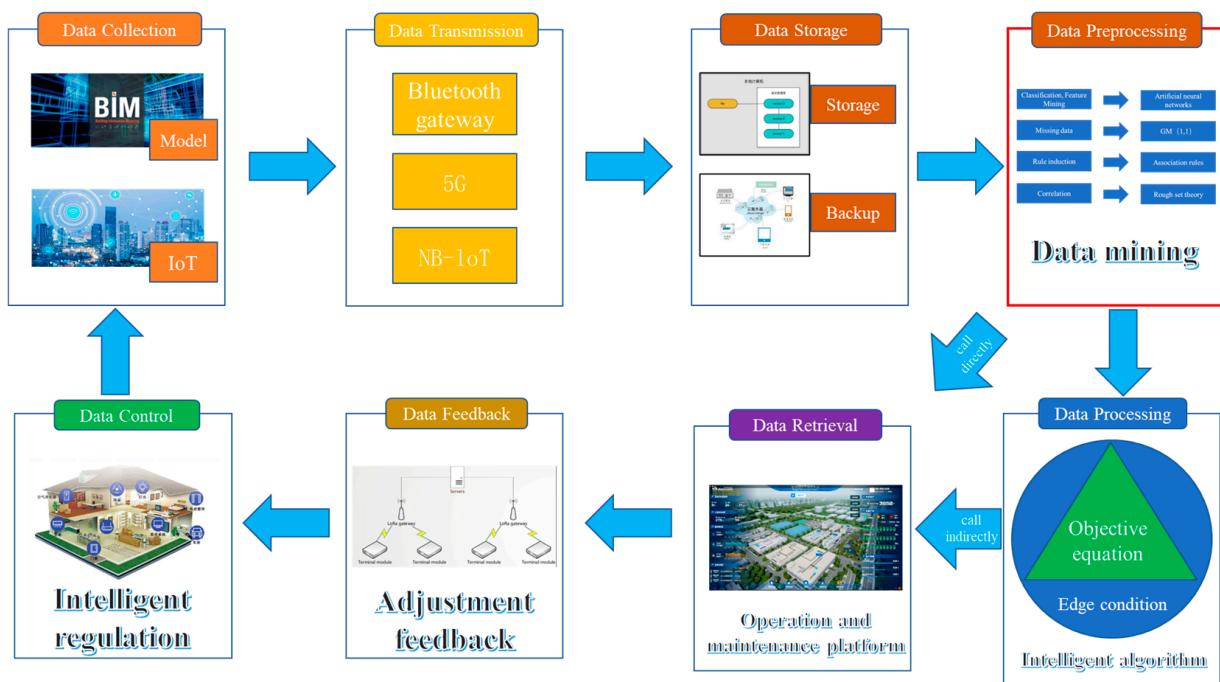


Figure 1. Data processing flow.

The main research focus of this study is centered on information preprocessing to ensure data completeness and precision. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format [11]. In order to guarantee the authenticity and reliability of the data, and to make them applicable for dynamic energy regulation and energy-saving strategy research in zero-carbon buildings, error adjustments were applied to energy-related data, catering to the needs of building operation and maintenance.

A comprehensive database model is established for supporting the operation and maintenance of zero-carbon buildings, building upon the foundation of the data preprocessing model. This structured database is effectively divided into four key categories, namely, building information [14–16], personnel and environmental data [5,17,18], energy consumption records [14,15], and functional equipment details (Figure 2). Architectural information constitutes an inherent attribute of a building, encompassing details such as the main body of the building and its geographical location. These attributes are typically resistant to change. In contrast, personnel and environmental information represent variables that constantly evolve, constituting uncertain factors within the building. However, they are closely tied to the user experience, particularly in intelligently adjustable zero-carbon buildings, where these factors play a pivotal role in autonomous regulation. The targets of such regulation include energy collection equipment and functional devices. The ability to collect energy consumption data stands out as one of the most significant features distinguishing zero-carbon buildings from conventional structures, making it a key element in their transformation into zero-carbon buildings. Functional equipment information pertains to data related to the usage of various equipment during the normal operation of zero-carbon buildings. Upon delving deeper into the research, it becomes evident that certain shared attributes exist among the categorized data. Notable examples include the correlation between factors like sunlight duration and solar panel power generation, as well as insights into energy consumption for lighting and the duration of operation under brightness levels below 15 lux. Leveraging these interconnected relationships, our investigation extends into the realm of data mining methodologies, enabling us to extract meaningful insights from the data-rich environment.

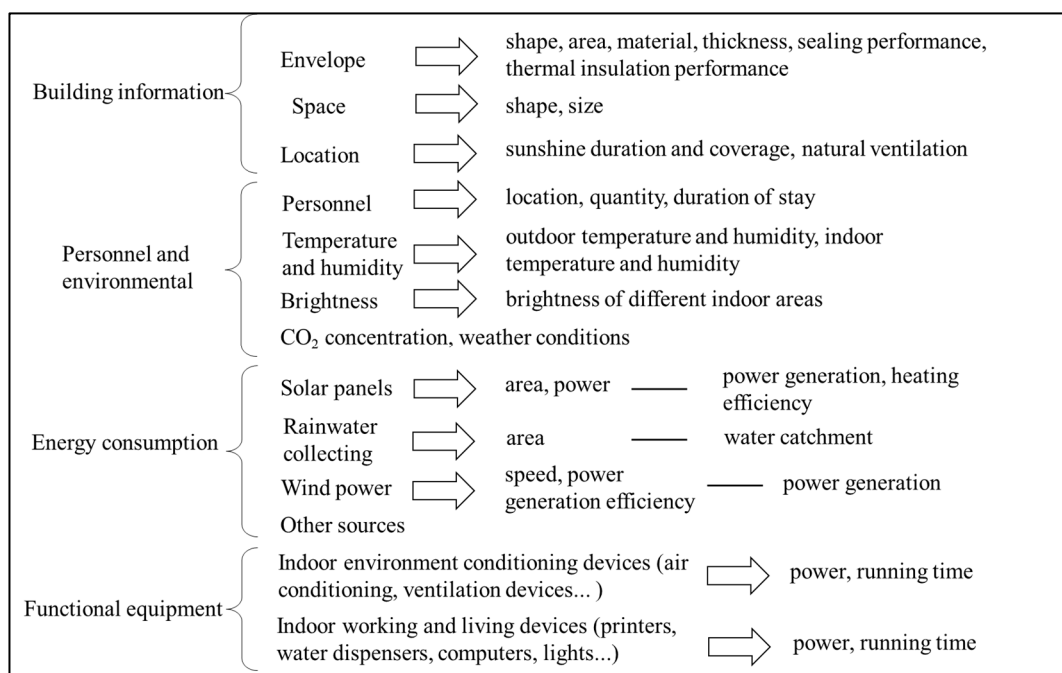
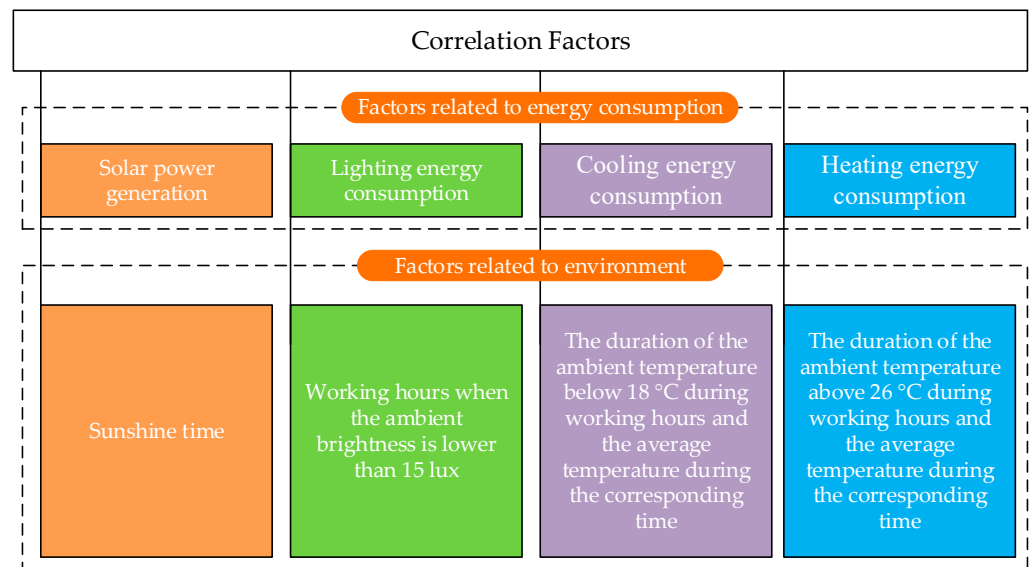


Figure 2. Database model.

### 3. Method

#### 3.1. Data Refining

Through a thorough exploration of data pertinent to zero-carbon buildings, intriguing relationships between collected energy consumption, utilization patterns, and building–environmental attributes have come to light. This discovery has paved the way for addressing optimization challenges at the data level. Building upon these insightful conjectures and informed by practical experiences, the pertinent information has been meticulously refined (Figure 3). Our refinement process has homed in on two core aspects: energy-related factors and environmental influences. This dual focus stems from the understanding that, within the context of minimizing human-driven variables, energy consumption predominantly derives from and is regulated by the surrounding environment. This intricate interplay becomes particularly significant for intelligent zero-carbon buildings that possess the capability to autonomously adapt [19], mitigating potential waste scenarios and fostering optimal resource utilization.



**Figure 3.** Analysis of related factors.

Among some related elements, there are linear correlations, such as between solar power generation and sunshine time, and between lighting energy consumption and working time when the ambient brightness is lower than 15 lux. At the same time, it was found that there is a trend correlation among other non-linear correlation factors. Simply put, there is a certain monotonic relationship between two elements, such as cooling energy consumption and heating energy consumption and their corresponding environmental factors. The research on data anomaly detection is carried out based on these two types of correlations.

#### 3.2. Error Discrimination Based on Correlation Test

To begin, correlation discovery is initiated based on the calculated data, with the aim of identifying standardized correlations between pairs of factors, denoted as  $C_0$ . This value is determined through multiple random calculations, representing the correlated data pairs of energy consumption and environmental factors over multiple time intervals. After analyzing the fluctuation range of correlations between these two factor types, we establish a constant value as the standard correlation, typically the average of correlations from random calculations.

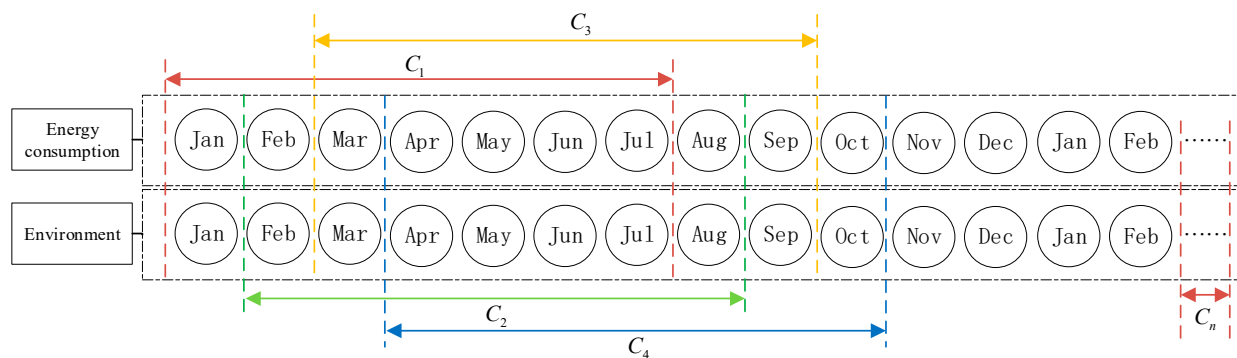
Subsequently, using  $C_0$  as the foundation, the error degree  $E$  for each dataset is calculated based on Equation (1). Since the range of calculations often fails to account for extreme values, the practical scope used for anomaly testing extends to 1.2 times the actual

calculation range. Once the range for  $E$  used in anomaly data testing has been established, it serves as a criterion for subsequent data anomaly checks. Data falling within this range are considered acceptable, while data falling outside this range are flagged as anomalous.

Figure 4 illustrates the specific steps involved in correlation computation. Each circle represents energy consumption and environmental information corresponding to individual months within a year. Continuous sets of seven months are considered as a testing step, and the corresponding correlations between data pairs for these seven months are calculated. The correlation for the first set of seven months is denoted as  $C_1$ , and this pattern continues for subsequent sets, labeled as  $C_n$ . Equation (1) is employed to compute the corresponding error degree, and this error is used to determine whether the data are anomalous or not.

$$E = \left| \frac{C_0 - C_n}{C_0} \right| \quad (1)$$

where  $E$  is the degree of error and  $C_n$  is the correlation of the corresponding data of the  $n$ th group of seven months.



**Figure 4.** Correlation calculation steps.

When the value of  $E$  is less than or equal to the upper limit of the abnormal inspection range, the data are considered to adhere to the correlation law and are deemed as normal data. Conversely, when the value of  $E$  exceeds the upper limit of the abnormal inspection range, the data are identified as abnormal and a fitting function is applied for correction. The standard correlations used to measure the upper limit of the abnormal test range for  $E$  can be categorized into two types: linear correlation factors and trend correlation factors. Linear correlation factors represent cases where two variables exhibit a clear linear relationship and can be analyzed through Pearson correlation testing. In this scenario,  $a_n$  represents the standard correlation for the  $n$ th type of linear correlation factors, equivalent to  $C_0$  in the Equation (1). Trend-related factors refer to situations where two variables exhibit a clear monotonic relationship but not a linear one and can be analyzed through Spearman correlation testing. In this case,  $b_n$  represents the standard correlation for the  $n$ th type of trend-related factors, also equivalent to  $C_0$  in Equation (1). It is important to note that the values of  $a_n$  and  $b_n$  may vary across different buildings, thus requiring recalculation for each building.

### 3.3. Error Correction and Test

On one hand, the limited volume of available data inherently poses challenges, potentially leading to issues such as under-fitting, over-fitting, and prediction model instability when employing the learning and predictive functionalities of neural network models for data correction. On the other hand, upon close examination of the existing correlation factors, it becomes evident that these amalgamated factors predominantly fall into two distinct categories: linear correlation factors and trend correlation factors. Linear correlation factors embody instances where a clear linear relationship exists between two types of factors. In contrast, trend correlation factors encompass pairs of factors that exhibit substantial

positive or negative correlation, even though they might not be strictly linearly related. Considering this distinction, a simplified model can be adeptly tailored to suit this type of data. Techniques such as linear fitting and polynomial fitting prove to be well-suited for effectively accommodating these characteristics.

The concept of the linear correlation factor revolves around harnessing its inherent linear relationship to rectify errors. This approach is particularly effective for correlation factors characterized by linear associations, such as lighting energy consumption and lighting duration, as well as solar power generation and sunshine duration. This correction technique is facilitated through a fitting formula derived from Equation (2) [20]. Optimizing the parameter solution for linear fitting involves the utilization of the least squares method. This method seeks to identify the most suitable fitting parameters by minimizing the sum of squared errors between actual observations and the theoretical model, represented by the fitting function. Let us consider our fitting function as  $y_i^* = f(x_i, \theta)$ , where  $\theta$  symbolizes the fitting parameter. The underlying principle of the least squares method is to determine the optimal fitting parameter  $\theta$  that minimizes the sum of squared errors, denoted as  $E(\theta)$  in Equation (3). This summation encapsulates the discrepancies between the observed values  $y_i$  and the anticipated values obtained from the predictive model  $f(x_i, \theta)$ .

$$y = ax + b \quad (2)$$

where  $x$  is the independent variable,  $y$  is the dependent variable,  $a$  is the slope of the line, and  $b$  is the intercept of the line.

$$E(\theta) = \sum (y_i - f(x_i, \theta))^2 \quad (3)$$

where  $y_i(1, 2, 3 \dots)$  is the actual observed value,  $f(x_i, \theta)$  is the predicted value, and  $\theta$  is the fitting parameter.

Trend-related factors predominantly encompass aspects related to heating and cooling. Through error discrimination grounded in correlation testing, it becomes evident that energy consumption and its corresponding duration do not conform entirely to linear correlation. Analysis indicates that this deviation is primarily due to the fact that heating and cooling energy consumption is influenced not only by the associated heating and cooling time but also by the temperature's fluctuation range. Upon delving into energy consumption itself, a distinct correlation with monthly trends emerges, illustrated by a parabolic scatter plot. This observation suggests the presence of a polynomial relationship [21]. Substantiated by the case study's results, the existence of this polynomial relationship is affirmed. Given these attributes, polynomial fitting is opted for as the suitable method for carrying out the requisite error correction. This approach is rooted in Equation (4). The Levenberg–Marquardt algorithm is employed to determine the optimal fitting parameters [22]. Widely utilized for solving nonlinear fitting problems, this algorithm is a commonly employed nonlinear least squares optimization technique. It amalgamates the steepest descent method with the Gauss–Newton method to expedite the search for the best-fit solution. During the fitting process, the algorithm computes the residuals—the discrepancies between actual data points and the polynomial function. By dynamically adjusting the step size, direction of fitted parameters, and step size weights based on these residuals, the algorithm navigates the parameter space. Its objective is to uncover the paramount combination of fitting parameters that minimizes the sum of squared residuals. Consequently, the fitting function aligns more closely with actual data, thus minimizing fitting errors. This algorithm proves efficient in handling intricate nonlinear fitting problems, including the complexities associated with polynomial fitting.

The LM algorithm updates the parameter  $\theta$  using Equation (5). When the value of  $\lambda_k$  is small [23], the LM algorithm exhibits behavior akin to the gradient descent method. This characteristic enables it to converge more swiftly within flat regions of the parameter space. On the other hand, when  $\lambda_k$  takes a larger value, the LM algorithm mirrors the behavior of the Gauss–Newton method, making it adept at accommodating steep regions in the

parameter space. Consequently, the LM algorithm dynamically fine-tunes the value of  $\lambda_k$  based on the behavior of the residuals during each iteration. This adaptive adjustment of  $\lambda_k$  ensures a more resilient optimization process.

$$y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n \quad (4)$$

where  $x^n$  ( $n = 1, 2, 3 \dots$ ) is the power of the independent variable  $x$ ,  $y$  is the dependent variable, and  $a_n$  is the fitting parameter.

$$\theta_{k+1} = \theta_k - (J_k^T J_k + \lambda_k I)^{-1} J_k^T r_k \quad (5)$$

where  $\theta_k$  represents the parameter vector at step  $k$ ,  $J_k$  signifies the Jacobian matrix (partial derivative matrix) of the fitting function  $P(x; \theta)$  concerning the parameter  $\theta$ ,  $r_k$  denotes the residual vector,  $\lambda_k$  is the parameter governing the control of step size, and  $I$  stands for the identity matrix.

This method primarily accomplishes data mining and preprocessing for zero-carbon buildings through error identification and correction (Figure 5). The data associated with energy consumption factors and their corresponding environmental factors are extracted from the database, and the nature of their data relationships is initially determined. The abnormal data are identified using the correlation test-based error detection method. If they are determined to be abnormal data, and the factor correlation type is linear, error correction is performed through linear fitting; if the factor correlation type is trending, polynomial fitting is applied for error correction. Subsequently, after error correction, another abnormality assessment is conducted, and the correction results are fed back into the database if the criteria for normality are met.

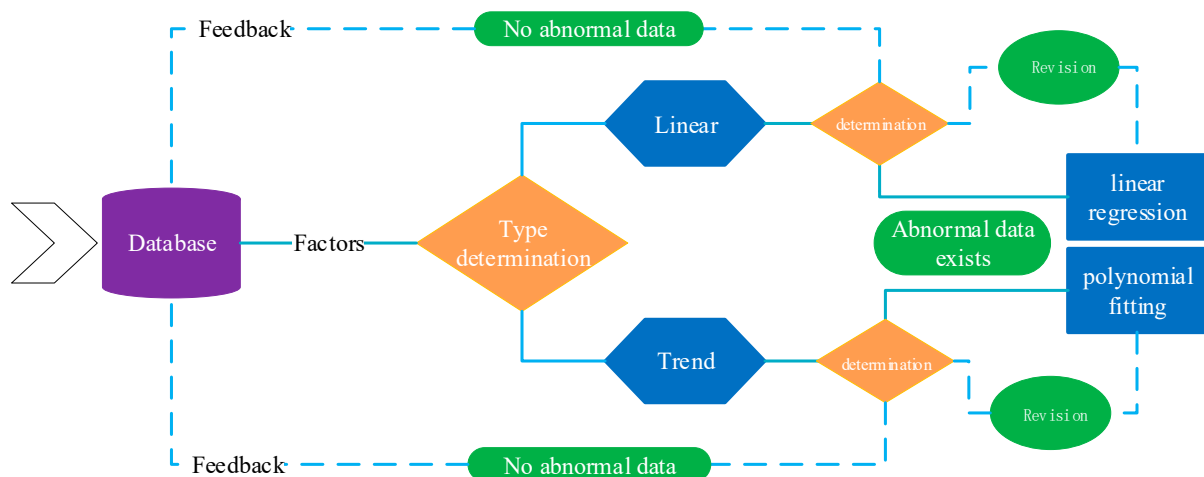


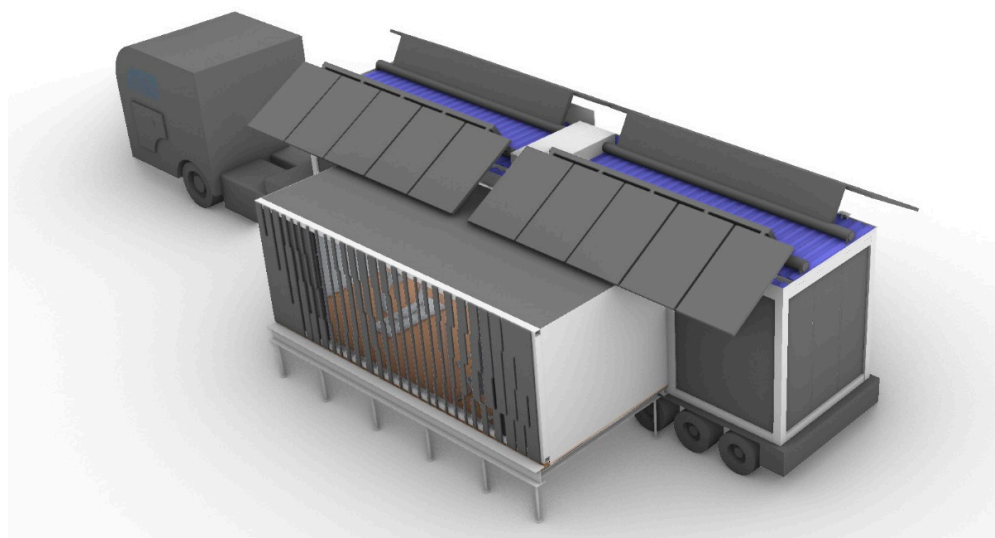
Figure 5. Flow steps of error discrimination, correction, and inspection.

#### 4. Case Study

This case study originates from an experimental zero-carbon initiative situated on the premises of Beijing University of Technology—an endeavor known as the “zero-carbon hut” (Figure 6). Spanning an area of approximately 50 m<sup>2</sup> with a total building height of 3 m, this hut comprises three distinct spaces: the power equipment room, the water system equipment room, and the normal activity room. It includes specific features such as rotatable solar panels with an embedded circulating water cooling system, rainwater collection equipment with a water treatment system, dual energy storage cabinets and power supply systems, scalable spaces, a running-based power generation device, and a sensor-linked digital twin control system. The fundamental objective of this project is to attain carbon neutrality. This is accomplished by leveraging high-precision sensing



technology in conjunction with an intelligent display platform to dynamically regulate the energy consumption throughout the hut.



**Figure 6.** BIM model of the zero-carbon hut.

This project endeavors to realize the concept of achieving carbon neutrality through a fully intelligent approach for cottage operations and maintenance. For instance, the illumination system is designed to assess occupancy within the cottage as well as the ambient light levels. The heating and cooling strategy is fine-tuned to maintain temperatures within the range of 18 °C to 26 °C. Furthermore, the water treatment system's efficiency is contingent on the number of occupants within the cottage. Drawing from these unique characteristics, our study focuses on exploring data mining methodologies specifically tailored to intelligent zero-carbon buildings.

During the data monitoring process, it has come to light that certain discrepancies arise in the monitored data due to factors such as sensor accuracy and external environmental challenges. These occasional inaccuracies in the collected data can subsequently lead to unforeseen impacts on the dynamic regulation process. As a countermeasure, it becomes imperative to subject the data to a screening and fine-tuning procedure to enhance the integrity of the data structure. In light of the pertinent factors illustrated in Figure 3, a collection of relevant characteristic data has been gathered. This data compilation serves as the foundation for refining our understanding and establishing a robust framework for subsequent analysis and optimization.

The perspective is held that, during the course of building operations, inherent errors exist within the correlation of two ostensibly interconnected data types. These errors are acknowledged and incorporated into the considerations. However, when these errors deviate beyond their original scope, it becomes essential to devise an analytical methodology aimed at rectifying these deviations. This measure is crucial to mitigate any potential disruptions to the building's operational and maintenance strategies. It is noteworthy that certain datasets exhibit notably positive correlations, such as the relationship between solar energy collection's energy consumption and the duration of sun exposure. Similarly, there is a correlation between daylight energy consumption and the duration of brightness below 15 lux during working hours. However, there are still some data that show a trend correlation. For instance, there is a correlation between refrigeration energy consumption and the duration of temperature falling below 18 °C during working hours, along with the average temperature within the corresponding timeframe (calculated as the cumulative sum of hourly temperatures divided by the corresponding hours). This correlation follows a monotonic trend.

It is essential to highlight that, given the project's limited operational timeframe, efforts were directed towards amassing a comprehensive dataset to substantiate the data mining methodologies outlined in this paper. To this end, an energy consumption model for the zero-carbon hut was constructed using EnergyPlus. The subsequent validation of the data mining techniques was conducted via the analysis of simulation results. To maintain data consistency, it is pertinent to note that all the data employed for case verification were exclusively simulated and did not involve real-time measurements. This approach was adopted to ensure the accuracy and reliability of the validation process.

#### 4.1. Error Discrimination and Data Mining of Linear Correlation Factors

As depicted in Table 1, the provided data show the factors associated with monthly lighting energy consumption and solar power generation for the zero-carbon hut. These factors have been derived through EnergyPlus simulation utilizing the EPW file corresponding to a standard year.

**Table 1.** Lighting energy consumption- and solar power generation-related factors.

| Month | Lighting Energy Consumption (kwh/m <sup>2</sup> ) | Lighting Duration (h) | Solar Power Generation (kwh) | Sunshine Duration (h) |
|-------|---|-----------------------|------------------------------|-----------------------|
| 1     | 0.176   | 166                   | 606.615                      | 140                   |
| 2     | 0.160   | 152                   | 715.56                       | 128                   |
| 3     | 0.151   | 147                   | 1099.76                      | 221                   |
| 4     | 0.145   | 139                   | 1336.359                     | 260                   |
| 5     | 0.151   | 144                   | 1569.64                      | 316                   |
| 6     | 0.120   | 111                   | 1526.885                     | 305                   |
| 7     | 0.124   | 114                   | 1377.915                     | 280                   |
| 8     | 0.124   | 116                   | 1269.254                     | 255                   |
| 9     | 0.146   | 140                   | 1194.476                     | 238                   |
| 10    | 0.150   | 143                   | 883.989                      | 200                   |
| 11    | 0.146   | 142                   | 587.012                      | 130                   |
| 12    | 0.176   | 166                   | 530.197                      | 110                   |

Before embarking on data mining efforts, simulated measurements on the pertinent data are initiated. By conducting iterative measurements across multiple datasets, foundational correlations are established. For instance, consider the correlation between lighting energy consumption and lighting duration, resulting in a coefficient ( $a_1$ ) of 0.99608, accompanied by an associated error range of [0, 0.004). Similarly, the correlation between solar power generation and sunshine duration yields a coefficient ( $a_2$ ) of 0.99078, with a corresponding error range of [0, 0.01). Additionally, the utilization of Pearson correlation analysis enables the computation of error degrees for linear correlation factors, as elucidated in Table 2.

Given that this dataset is a product of simulation, the values are derived through application of the pertinent energy consumption theory. This data collection adheres to a relatively standardized methodology. The level of error in the "lighting energy consumption" and "solar power generation" correlations adheres to the prescribed criteria, rendering any correction unnecessary. Furthermore, this alignment underscores the feasibility and applicability of the underlying theory.

Given the linear correlation attributes exhibited by the data, employing linear regression for data correction proves suitable. Below is presented the linear regression equation alongside the corresponding residual diagram (Figure 7). Evidently, the depicted figures reveal a fitting  $R^2$  value of 0.9848 for lighting-related factors and 0.9798 for solar power generation-related factors. As observed through the distribution of corresponding residuals, a uniform distribution prevails, meeting the criteria for normal distribution. This substantiates the effectiveness of the fitting approach, rendering it a viable equation for error correction purposes.

Table 2. Error calculation.

| Month | Lighting Energy Consumption and Lighting Time | Degree of Error | Solar Power Generation and Sunshine Duration | Degree of Error |
|-------|---|-----------------|--|-----------------|
| 1–7   | 0.99393                                       | C <sub>11</sub> | 0.99074                                      | C <sub>21</sub> |
| 2–8   | 0.99602                                       | C <sub>12</sub> | 0.99731                                      | C <sub>22</sub> |
| 3–9   | 0.99773                                       | C <sub>13</sub> | 0.99446                                      | C <sub>23</sub> |
| 4–10  | 0.99842                                       | C <sub>14</sub> | 0.98532                                      | C <sub>24</sub> |
| 5–11  | 0.99634                                       | C <sub>15</sub> | 0.99555                                      | C <sub>25</sub> |
| 6–12  | 0.99403                                       | C <sub>16</sub> | 0.99486                                      | C <sub>26</sub> |

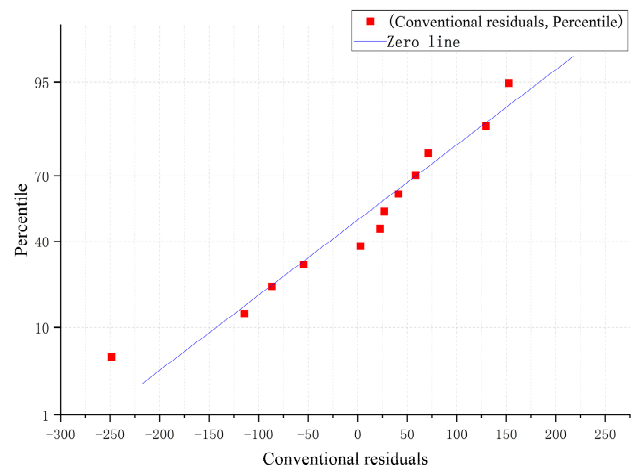
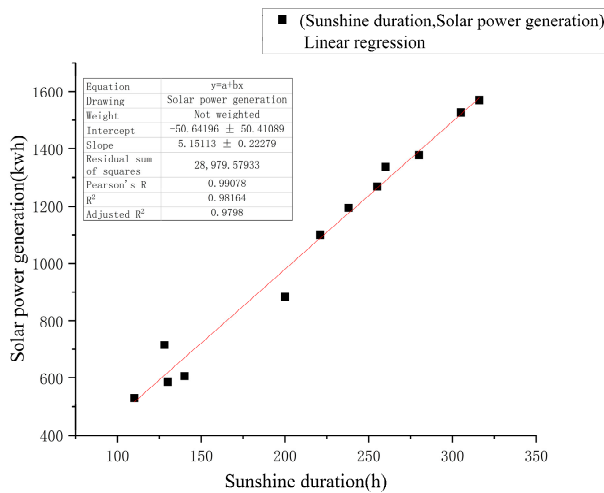
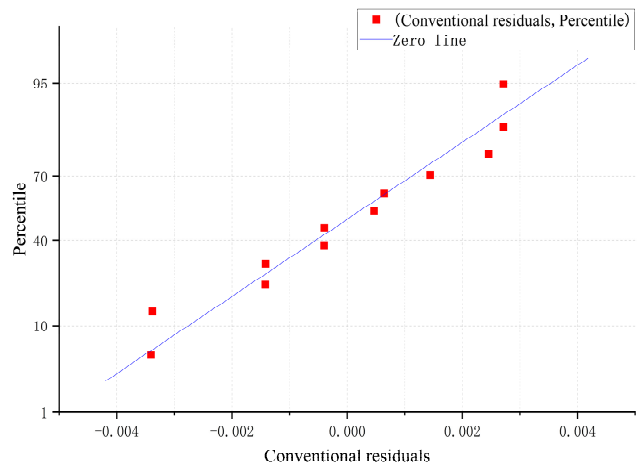
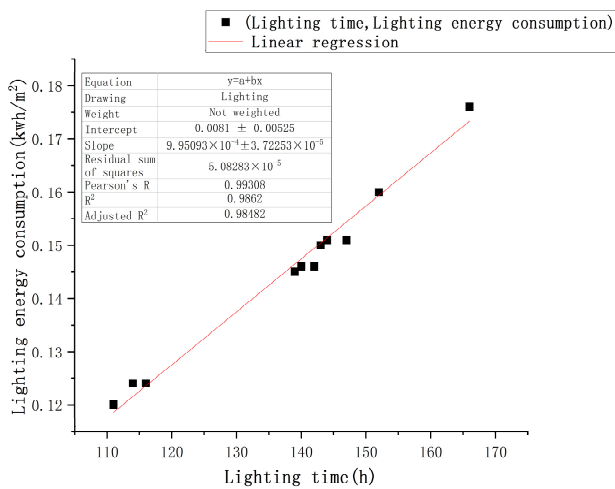


Figure 7. Regression equation and corresponding residual distribution of factors related to lighting and solar power generation.

4.2. Error Discrimination and Data Mining of Trend Correlation Factors

The data table reflects the pertinent elements of cooling and heating energy consumption for the zero-carbon hut (Table 3). These data are derived from EnergyPlus simulations utilizing the EPW file for a standard year.

**Table 3.** Related factors of cooling and heating energy consumption.

| Month | Heating Energy Consumption (kwh/m <sup>2</sup> ) | Working Hours below 18 °C (h) | Average Temperature within the Time Range of Less than 18 during Working Hours (°C) | Cooling Energy Consumption (kwh/m <sup>2</sup> ) | The Duration of Working Time Higher than 26 °C (h) | Average Temperature in the Time Range above 26 °C during the Working Time (°C) |
|-------|--|-------------------------------|---|--|--|--|
| 1     | 6.641  | 480                           | −2.74   | 0  | 0  | 26   |
| 2     | 3.738  | 420                           | 1.42  | $3.307 \times 10^{-7}$                           | 0  | 26   |
| 3     | 1.604  | 353                           | 6.43  | 0.0003   | 0  | 26   |
| 4     | 0.065  | 241                           | 13.027  | 0.446  | 18   | 28.789   |
| 5     | 0.032  | 81                            | 14.993  | 2.328  | 124  | 28.426   |
| 6     | 0  | 5                             | 17.8  | 4.414  | 237  | 29.230   |
| 7     | 0  | 0                             | 18  | 6.012  | 300  | 29.087   |
| 8     | 0  | 5                             | 17.6  | 5.026  | 211  | 28.59  |
| 9     | 0.013  | 61                            | 15.554  | 2.258  | 60   | 26.857   |
| 10    | 0.181  | 384                           | 12.572  | 0.188  | 4  | 26.5   |
| 11    | 1.913  | 447                           | 6.374   | $1.0235 \times 10^{-6}$                          | 0  | 26   |
| 12    | 4.842  | 465                           | 0.454   | 0  | 0  | 26   |

Before commencing data mining efforts, preliminary simulated measurements are conducted on the relevant data. By repeatedly measuring multiple sets of data, foundational correlations are established. For instance, consider the correlation between heating energy consumption and equipment heating duration, yielding a coefficient ( $b_1$ ) of 0.9806134, accompanied by an associated error range of [0, 0.02). Similarly, the correlation between cooling energy consumption and equipment cooling duration results in a coefficient ( $b_2$ ) of 0.9588279, with a corresponding error range of [0, 0.01). This was combined with Spearman correlation analysis to calculate the error degree of trend-related factors (Table 4).

**Table 4.** Spearman correlation analysis.

| Month | Heating Energy Consumption and Equipment Heating Time | Degree of Error | Refrigeration Energy Consumption and Equipment Cooling Time | Degree of Error |
|-------|---|-----------------|---|-----------------|
| 1–7   | 0.9910312   | $C_{11}$        | 0.9636241   | $C_{21}$        |
| 2–8   | 0.9723449   | $C_{12}$        | 0.9549937   | $C_{22}$        |
| 3–9   | 0.9723449   | $C_{13}$        | 0.9642857   | $C_{23}$        |
| 4–10  | 0.9723449   | $C_{14}$        | 0.9642857   | $C_{24}$        |
| 5–11  | 0.9723449   | $C_{15}$        | 0.9642857   | $C_{25}$        |
| 6–12  | 0.9723449   | $C_{16}$        | 0.9549937   | $C_{26}$        |

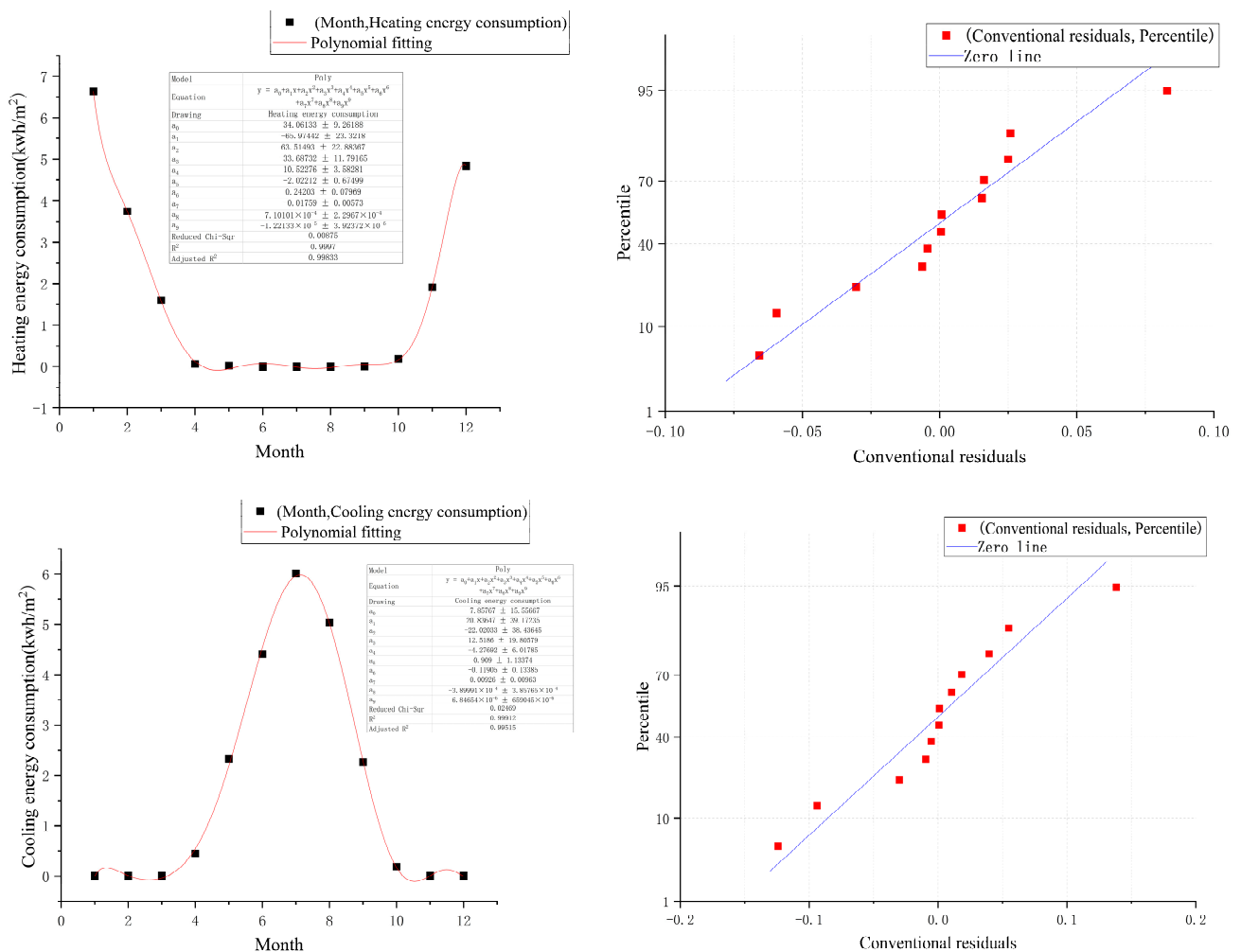
Based on the test results, the error degrees are generally in line with the prescribed criteria. To further substantiate the data's rationality, efforts were extended through logical examination. A distinctive feature here lies in the collection of average temperatures for heating energy consumption and refrigeration energy consumption. Unlike other data types, the average temperature is calculated from hourly temperature monitoring data extracted from the EPW file, combined with equipment operational hours. Consequently, while the accuracy of these data might not meet the prerequisites for a Spearman correlation test, it remains suitable for logical validation. As presented in Table 5, a negative correlation ranging between −0.95 and −0.8 is observed between heating energy consumption and the average temperature within the heating time frame. Conversely, a positive correlation ranging between 0.6 and 0.9 emerges between refrigeration energy consumption and the average temperature within the cooling time frame. This logical alignment aligns seamlessly with typical data cognition, thereby affirming its conformity with the relevant requirements.

**Table 5.** Spearman correlation analysis.

| Month | Heating Energy Consumption and Average Temperature within the Heating Time Range | Cooling Energy Consumption and Average Temperature within the Cooling Time Range |
|-------|--|--|
| 1–7   | −0.9636241   | 0.8894992  |
| 2–8   | −0.9063270   | 0.7387687  |
| 3–9   | −0.8017837   | 0.6428571  |
| 4–10  | −0.8017837   | 0.6428571  |
| 5–11  | −0.8017837   | 0.8928571  |
| 6–12  | −0.9063270   | 0.8829187  |

Since the data being utilized originate from simulations, the capability to verify the method’s accuracy is limited without undergoing a process of cross-referencing inaccurate data. Therefore, in the conclusion of the methodology only, a corrective measure for abnormal data is provided: the employment of Polynomial fitting for error rectification.

The fitting results (Figure 8) demonstrate that the  $R^2$  values for the fitting functions concerning cooling energy consumption and heating energy consumption attain 0.9983 and 0.9951, respectively. Examining the distribution of the corresponding residuals reveals a uniformly distributed pattern that adheres to the criteria of a normal distribution. Consequently, this fitting function emerges as notably effective, serving as a solid foundation for error correction.



**Figure 8.** Regression equations and corresponding residual distribution of heating and cooling related factors.

## 5. Results and Discussion

This method primarily addresses precision issues stemming from various error factors in the measurement of data related to zero-carbon smart building operation and maintenance. Unlike most data anomaly detection and correction methods, this approach seamlessly combines theoretical principles with domain knowledge. However, it does not demand an extensive depth of expertise in either theoretical methods or domain knowledge, making it more versatile. This method does not involve the application of complex forecasting algorithms [3–6] or require a strong theoretical development background [7–11]. Its core lies in the calculation of the error degree “ $E$ ” in Equation (1), the consideration of energy consumption–environment related factors (Figure 3), and the correlation calculation step (Figure 4). The synergy of these three components simplifies and streamlines data mining and preprocessing, endowing it with robust explanatory capabilities.

Judging from its application effect on the case of the zero-carbon hut, this method can be effectively adapted to associated data. There are corresponding detection methods for various types of building energy consumption, and data mining and forecasting are carried out on a monthly basis. Considering the limited number of data, this method establishes appropriate correction functions tailored to the distinctive characteristics of different data types. Based on the results, the correction functions have achieved significant precision, with a coefficient of determination ( $R^2$ ) exceeding 0.99, outperforming the majority of specially developed energy consumption prediction models.

## 6. Conclusions

This paper primarily delves into the preprocessing analysis of operational and maintenance data from zero-carbon smart buildings. Through a meticulous examination of the correlation between energy consumption data and environmental factors, the data’s validity is confirmed, aligning with the requisites of subsequent operational and maintenance phases. The study primarily focuses on four key aspects: lighting energy consumption and lighting duration, solar power generation and effective sunshine duration, heating energy consumption and equipment heating operational hours, and cooling energy consumption and equipment cooling operational hours.

It is worth noting that the specificities of distinct cases may lead to variations in the standard correlation  $C_0$  between corresponding energy consumption and environmental factors. Moreover, since the case data stem from simulations, the presence of simulated abnormal data is absent. However, during real-world case verification, if the error range criteria are not met, the establishment of fitting functions can be explored to effect further adjustments, thus aligning with the necessary standards.

The primary focus of this method is to address errors that occur during the data collection process of zero-carbon building energy consumption, encompassing tasks like outlier removal and missing value interpolation. In the future, the widespread adoption of data augmentation, transfer learning, and semi-supervised learning techniques for handling extensive datasets related to building operations is anticipated [24]. However, challenges are still presented by these methods, such as a heavy reliance on theory and limited interpretability. The research findings presented in this paper effectively mitigate these shortcomings and serve as valuable references within the realm of current preprocessing technologies. This contribution has the potential to foster advancements in data-driven research within the field of zero-carbon buildings.

Additionally, the approach substantiates the data’s rationality through correlation analysis, anchoring it on the relationship between energy consumption factors and environmental factors. The process is characterized by its simplicity and practicality, rendering it apt for intelligent operation and facilitating data management across diverse zero-carbon building scenarios. Ultimately, this method contributes valuable insights to the realm of the operation and maintenance of zero-carbon buildings. Furthermore, the case presented in this paper serves as a verification example, and its applicability can be expanded to encompass general buildings equipped with self-regulating capabilities, provided that all

pertinent factors are thoroughly explored and considered. Certainly, it can be stated that, in the future, the more comprehensively the relationship between energy consumption factors and environmental factors within buildings is explored, the broader the applicability of this approach will become.

**Author Contributions:** Conceptualizing, review and editing, Y.Z.; Data collection, writing, drafting-Original draft, R.L.; Supervision, review and funding acquisition, Z.L.; resources and review, Y.L.; Supervision, review and editing, L.L.; Supervision, J.W.; Data collection, W.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

- Chen, L.; Msigwa, G.; Yang, M.; Osman, A.I.; Fawzy, S.; Rooney, D.W.; Yap, P. Strategies to achieve a carbon neutral society: A review. *Environ. Chem. Lett.* **2022**, *20*, 2277–2310. [[CrossRef](#)] [[PubMed](#)]
- Chen, L.; Huang, L.; Hua, J.; Chen, Z.; Wei, L.; Osman, A.I.; Fawzy, S.; Rooney, D.W.; Dong, L.; Yap, P. Green construction for low-carbon cities: A review. *Environ. Chem. Lett.* **2023**, *21*, 1627–1657. [[CrossRef](#)]
- Fu, Y.; Li, Z.; Feng, F.; Xu, P. Data-quality detection and recovery for building energy management and control systems: Case study on submetering. *Sci. Technol. Built Environ.* **2016**, *22*, 798–809. [[CrossRef](#)]
- Ma, L.; Huang, Y.; Zhao, T. A synchronous prediction method for hourly energy consumption of abnormal monitoring branch based on the data-driven. *Energy Build.* **2022**, *260*, 111940. [[CrossRef](#)]
- Lundstrom, L. Adaptive weather correction of energy consumption data. In Proceedings of the 8th International Conference on Applied Energy (ICAE2016), Beijing, China, 8–11 October 2016; Beijing Inst Technol: Beijing, China, 2017; pp. 3397–3402.
- Alghamdi, T.A.; Javaid, N. A Survey of Preprocessing Methods Used for Analysis of Big Data Originated from Smart Grids. *IEEE Access* **2022**, *10*, 29149–29171. [[CrossRef](#)]
- Chen, L.; Lao, K.; Ma, Y.; Zhang, Z. Error Modeling and Anomaly Detection of Smart Electricity Meter Using TSVD+L Method. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3201940. [[CrossRef](#)]
- Choi, J.; Lee, I.; Cha, S. Cluster Analysis to Preprocess the Building Power Usage Data Without Domain Knowledge. *J. Electr. Eng. Technol.* **2020**, *15*, 685–692. [[CrossRef](#)]
- Zach, R.; Hofstaetter, H.; Mahdavi, A. A distributed and scalable approach to building monitoring. In *Ework and Ebusiness in Architecture, Engineering and Construction 2014*; Department of Building Physics and Building Ecology of the Vienna University: Vienna, Austria, 2015; pp. 11–18.
- Tauber, C.; Tahmasebi, F.; Zach, R.; Mahdavi, A. Automated simulation model calibration based on runtime building monitoring. In *Ework and Ebusiness in Architecture, Engineering and Construction 2014*; Department of Building Physics and Building Ecology of the Vienna University: Vienna, Austria, 2015; pp. 265–270.
- Bhagat, V.; Robins, B.; Pallavi, M.O. Sparx—Data preprocessing module. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Pune, India, 29–31 March 2019.
- Shen, K.; Ding, L.; Wang, C.C. Development of a Framework to Support Whole-Life-Cycle Net-Zero-Carbon Buildings through Integration of Building Information Modelling and Digital Twins. *Buildings* **2022**, *12*, 1747. [[CrossRef](#)]
- Olariu, E.M.; Tolas, R.; Portase, R.; Dinsoreanu, M.; Potolea, R. Modern approaches to preprocessing industrial data. In Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP 2020), Electr Network, Cluj-Napoca, Romania, 3–5 September 2020; pp. 221–226.
- Miller, W.; Buys, L. Anatomy of a sub-tropical Positive Energy Home (PEH). *Sol. Energy* **2012**, *86*, 231–241. [[CrossRef](#)]
- Mata, E.; Korpala, A.K.; Cheng, S.H.; Jimenez Navarro, J.P.; Filippidou, F.; Reyna, J.; Wang, R. A map of roadmaps for zero and low energy and carbon buildings worldwide. *Environ. Res. Lett.* **2020**, *15*, 113003. [[CrossRef](#)]
- Anderson, N.; Wedawatta, G.; Rathnayake, I.; Domingo, N.; Azizi, Z. Embodied Energy Consumption in the Residential Sector: A Case Study of Affordable Housing. *Sustainability* **2022**, *14*, 5051. [[CrossRef](#)]
- Qiao, J.; Zhang, X.; Hao, C.; Liu, S.; Zhang, Y.; Xing, K.; Yang, P. Post-occupancy evaluation of the actual performance of a low-carbon building. *Energy Rep.* **2023**, *10*, 228–243. [[CrossRef](#)]
- Trofimova, P.; Cheshmehzangi, A.; Deng, W.; Hancock, C. Post-Occupancy Evaluation of Indoor Air Quality and Thermal Performance in a Zero Carbon Building. *Sustainability* **2021**, *13*, 667. [[CrossRef](#)]
- Bhutta, F.M. Application of smart energy technologies in building sector—Future prospects. In Proceedings of the 2017 International Conference on Energy Conservation and Efficiency (ICECE), Lahore, Pakistan, 22–23 November 2017; pp. 7–10.

20. Kong, M.; Li, D.; Zhang, D. Research on the application of improved least square method in linear fitting. In Proceedings of the 2018 4th International Conference on Environmental Science and Material Application, Xian, China, 15–16 December 2019.
21. Torabi Moghadam, S.; Toniolo, J.; Mutani, G.; Lombardi, P. A GIS-statistical approach for assessing built environment energy use at urban scale. *Sustain. Cities Soc.* **2018**, *37*, 70–84. [[CrossRef](#)]
22. Murray, K.; Mueller, S.; Turlach, B.A. Revisiting fitting monotone polynomials to data. *Comput. Stat.* **2013**, *28*, 1989–2005. [[CrossRef](#)]
23. Rouhani, M.; Domingo Sappa, A. Implicit Polynomial Representation Through a Fast Fitting Error Estimation. *IEEE Trans. Image Process* **2012**, *21*, 2089–2098. [[CrossRef](#)] [[PubMed](#)]
24. Fan, C.; Chen, M.; Wang, X.; Wang, J.; Huang, B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data. *Front. Energy Res.* **2021**, *9*, 652801. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.