2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

ENHANCING DATA QUALITY VIA DATA CORRECTION TECHNIQUES TO ASSURE HIGH-QUALITY DATA FOR ENERGY SERVICES ACROSS EUROPE

José L. Hernández[1,2], Carlos de la Cruz[1], Ignacio de Miguel[2] and Fredy Vélez[2]
[1]CARTIF Technology Centre, Boecillo, Spain
[2]Universidad of Valladolid, Valladolid, Spain

Abstract

The growing adoption of digital technologies is reshaping how buildings are managed. Still, low data quality remains a key challenge. Incomplete or inaccurate data can compromise decisions and hinder compliance with regulations. The work presented here, tested on ten pilot buildings across Europe, significantly improves data quality, raising completeness from 55% to 100%, while improving accuracy. A centralized data lake ensures seamless integration of static and dynamic sources, enabling real-time insights. Beyond technical gains, the solution supports the goals of the Energy Performance of Buildings Directive, enhancing energy efficiency and sustainability through better data governance and scalable, replicable management models.

Introduction

Digital technologies continue rapidly evolving; therefore, transforming the construction and building management sectors. One of the biggest challenges in this transformation is the significant increase in the volume of data being generated. Tools like the Internet of Things (IoT), Artificial Intelligence (AI), Blockchain, Distributed Ledger Technology (DLT), and Big Data (BD) are now playing a pivotal role in reshaping how buildings are designed, operated, and maintained. At the same time, the integration of Building Automation and Control Networks (BACN), especially in commercial and tertiary buildings, is contributing to an ever-expanding data landscape and, in consequence, the volume of available data.

Despite this wave of digital innovation, the issue of data quality remains an important challenge and bottleneck (Morewood et al., 2023). Incomplete datasets, leading to poor data quality, can significantly undermine key processes in the building sector, affecting aspects such as energy performance assessments, predictive maintenance strategies, and financial planning in the built environment. For instance, in smart buildings, gaps or inaccuracies in sensor readings may lead to inefficient heating or cooling, driving up both energy use and operational costs. Moreover, the lack of a robust data governance framework hinders interoperability between systems, reducing the effectiveness of digital twins and automation solutions.

These challenges are particularly critical given the growing pressures in the European regulations, where directives such as the Energy Performance of Buildings Directive (EPBD) set higher standards for energy efficiency and carbon reduction. Inaccurate or unreliable data can make it difficult for building operators and decision-makers to meet these goals, potentially leading to misguided strategies and missed sustainability targets.

One of the most common data issues is the incomplete datasets, containing errors and having inconsistency, missing values or outliers. These aspects create obstacles for effective decision-making across the sector. Although many construction projects aim to harness data-driven methods for improved efficiency and process optimization, unreliable information often results in delays, increased risks, and reduced system functionality.

Particularly difficult is the management of the growing data flows, which leads to more complex traceability and error management in the vast streams of information generated daily (Hossein et al., 2020). Furthermore, it should be considered the large-scale datasets and the diversity of modern datasets, making it harder to maintain data quality, especially accuracy and consistency. In addition, communication disruptions often lead to technical faults, compromising the reliability of data sets and, thus, data-driven approaches (Yong et al., 2021).

Governance presents another major hurdle. Clear definitions around data ownership, especially for machine-generated data, are still lacking (Ender, 2021). In many cases, information is stored in disconnected silos, limiting interoperability and preventing the kind of integration necessary for advanced analytics (Abraham et al., 2019). Without a holistic approach and/or a cohesive data strategy, the sector risks missing out on valuable insights that could drive innovation and efficiency.

Beyond the technical implications, low data quality also impacts collaboration among stakeholders. Data interpretation can lead to misunderstandings, inconsistencies or contradictions. When project managers and teams lack access to reliable, up-to-date information, they may be forced to rely on assumptions, increasing

uncertainty and project risks. This often leads to errors in decision-making.

To unlock the full potential of digital transformation in the built environment, data quality should be prioritized, covering dimensions such as data completeness, accuracy, consistency, interoperability, governance, and integration. Without facing these challenges, the benefits of advanced ICTs and automation will remain limited, hindering progress toward smarter, more efficient, and resilient buildings.

This study addresses these challenges by introducing a methodology aimed at improving data quality. The approach combines structured evaluation techniques to, firstly, calculate data quality indicators, with machine learning tools to correct inconsistencies. The method focuses on mainly enhancing completeness, accuracy, and consistency in datasets used for smart building applications (Hernández, 2024), ensuring that datasets meet the required quality standards. By applying the solution to ten real pilot cases, the study demonstrates how data-driven strategies can significantly improve operational outcomes and help meet energy efficiency targets.

The rest of the paper is structured as follows. Next section describes a background about high-quality data in Smart Buildings. This is followed by the concept of the data management within this approach in order to continue with the methodological approach, providing some results about data quality analysis. Next, the data correction techniques are described to conclude with an example of application in one of the pilots.

Advancing towards high-quality data in Smart Buildings

The growing complexity of modern buildings and their management systems has amplified the need for reliable, high-quality data. Digitalization is transforming the built environment, with data playing a critical role in shaping how properties are utilized, monitored, and optimized. By monitoring energy use, air quality, and structural health, data provides valuable insights that help decision-makers improve efficiency. This reduces the cost in the building management and maintenance, creating better ways to control energy systems and, then, achieve sustainability targets.

Ensuring data accuracy, completeness, and consistency is essential for enhancing the potential in the smart building sector. Low data quality introduces risks that extend across the entire lifecycle of a building, from design and construction to operation and maintenance. Having higher data quality standard thanks to more transparent and reliable data processes not only support better informed decision-making, but also increased alignment with regulatory frameworks such as the Energy Performance of Buildings Directive (EPBD, 2023), which emphasizes energy efficiency, carbon footprint reduction, and occupant comfort.

In this context, high-quality data is defined by three key pillars (BDVA/DAIRO, 2021).

- Reliability & Credibility

Machine-learning algorithms can effectively improve and enhance data reliability. They are helpful to detect and correct outliers, ensuring that inconsistencies are minimized at every stage of data life cycle. This involves:

  - Addressing data gaps by developing interpolation methods to populate new data to complete missing dataset in time series. This also applies for static data, where inconsistencies are detected (e.g., digital logbooks and cadastral records).
  - Identifying and mitigating outliers. When out-of-range values are identified, uncertainty in the implementation of algorithms or decision-making processes distort the data-driven analysis, requiring contingency plan to mitigate these uncertainties.
  - Ensuring consistency and accuracy, when values are not following the normal distribution. Data duplicates are often detected in the data records. Also, synchronizing data from multiple sources, and standardizing ambiguous values are essential steps to improve data reliability and support effective analysis.
  - Validating digital models: In the case of Building Information Modelling (BIM), errors in static data often propagate throughout a building's lifecycle. BIM is used from design to operation; therefore, whether a BIM design error is introduced, operation strategies will be affected. That is why early detection of these inconsistencies can prevent inefficiencies in design and facility management.

- Interoperability

Heterogeneity of data in the building sector is very common as datasets are collected from multiple sources, including IoT sensors, energy performance databases, BIM platforms, and geographic information systems. Integrating all these datasets in a homogenized way remains as a challenge due to inconsistencies in format and structure. A standardized approach leveraging models such as NSGI-LD, SAREF, BRICK, and IFC ensures that data can be harmonized, improving its applicability across different use cases (Hernández, 2020).

- Privacy & Security

Quality assurance should also consider data privacy and security. Building sector involves particulars, which could deal with handling sensitive personal or operational information. Distributed Ledger Technology (DLT) and blockchain frameworks provide robust solutions for ensuring data integrity, privacy, and traceability, protecting both individuals and organizations from data breaches and unauthorized modifications.

The consequences of unreliable data extend beyond inefficiencies in energy management or real estate valuations. Inaccurate data has led to significant issues in real-world projects. For instance, infrastructure failures, where erroneous site survey data has resulted in faulty foundation designs, leading to costly structural instability. Another example is BIM coordination errors when

incomplete or inconsistent data inputs (Alsuhaibani et al., 2022) have caused clashes between different construction trades, leading to project delays (Koo et al., 2024).

Moreover, there are challenges in Energy Services (ESCO) models, such as inaccuracies in smart meter data collection complicate billing procedures and distort energy savings calculations, affecting financial planning and investment strategies (Chopade et al., 2024).

Addressing these challenges requires a holistic approach to data governance. That is why, as introduced, this paper presents a procedure that deals with the poor data quality samples by a two-step approach. First step consists of the data quality analysis (Hernández, 2024) to determine the real quality of the datasets. Secondly, correction techniques are applied so that quality of data can be improved to support high-quality Smart Building services. This second step is validated against one of the pilot buildings.

Data management approach

Data quality dimensions

Data quality is defined in different ways within the literature without a consensus in a common definition (Hernández, 2024b). Within this manuscript the definition from (Hernández, 2024) has been followed and it is summarized in Table 1.

Accuracy measures the percentage of data that falls within an expected range. It is important to remark the difference between "expected" and "measurement" range. The first one sets the usual values that the variable should take, like indoor temperature between 15-25ºC, while measurement range refers to the all possible values that a variable could span (e.g., a temperature probe could take values between -100 and 100ºC, but not possible for indoor temperature). Outliers, even if correctly measured, are identified to prevent their use in services.

Table 1: Data quality multi-dimension approach

Data quality dimension	Description of the dimension
Accuracy	Out of "expected" range detection
Completeness	Gap detection and interpolation
Reliability	Accuracy x completeness
Consistency	Detection of "normal" data patterns & Proof of Existence (PoE)
Relevance	Data filtering processes
Accessibility	Open APIs and protocols
Timeliness	Pre-analytics

Completeness evaluates data gaps by comparing the actual number of collected samples with the expected number based on sampling frequency. Missing data is quantified by the ratio between available and expected samples, ensuring datasets are as comprehensive as possible.

Reliability assesses whether data is feasible for analysis, service development, or sharing with third parties. It combines completeness and accuracy metrics to generate an overall quality score.

Consistency examines data patterns by analyzing their distribution around the mean and standard deviation. This helps ensure uniformity and detect deviations that may indicate inconsistencies.

Relevance determines which datasets are genuinely useful for end-user services. Since many data points may have no practical application, this metric helps filter out unnecessary information.

Accessibility measures how easily data can be retrieved, shared, and utilized, often relying on open protocols and APIs. However, some datasets may have restricted access, limiting their availability for broader applications.

Timeliness evaluates how up-to-date data is and measures the delay in gathering critical information. This metric is expressed in time units to define the acceptable maximum delay for decision-making.

Even though seven dimensions are defined, within this work, relevance, accessibility and timeliness are not considered. The reason lies in the fact that these dimensions depend on external factors, like use of the data in third party services, interfaces from field level to collect data or capabilities for real time of building management systems. Then, the application of data correction techniques would not be effective over the aforementioned dimensions.

Data management through a data lake concept

Data management relies on a data lake (Hernández, 2023b), which serves as a centralized storage system that retains vast amounts of raw data in its original format until required for processing. Unlike conventional databases, it accommodates structured, semi-structured, and unstructured data, offering greater flexibility in analysis. Several key aspects define its functionality:

- Decentralized approach by using local pilot buildings repositories (building management systems) rather than centralizing all data, instead focusing on extracting only relevant datasets based on data quality relevance criteria.
- Data synchronization of both dynamic and static repositories, which must be continuously aligned to enable meaningful knowledge extraction and support the creation of data marts.
- API-driven access via intelligent querying through APIs facilitates seamless data exchange with digital services and digital twins.

Figure 1 illustrates the data lake architecture, emphasizing the selective filtering of pilot data to ensure only relevant information is stored in the Data Warehouse (DWH).

These datasets are not only utilized for DWH population but also streamed in real-time via a Kafka broker, enabling digital twins to maintain up-to-date representations of pilot environments. Meanwhile, the API supports the sharing of historical and aggregated data, optimizing computational efficiency for machine learning and AI applications.

Additionally, static data, sourced through various channels such as BIM models, ETL metadata, or files (e.g., CSV, Excel) from pilot leaders, is synchronized with the DWH to maintain consistency (Hernández, 2023). Data marts extract information from both repositories, combining dynamic and static datasets to generate business-specific insights, such as thermal energy trends and aggregated analyses. The resulting datasets, referred to as enriched time-series data, serve as a valuable resource for advanced analytics and decision-making.

Data collection through the pilot buildings

Data collection from the pilot buildings is heterogeneous with multiple protocols and frequencies. Table 2 summarizes case-by-case data collection mechanisms.

Table 2: Pilot buildings within the EU DigiBUILD project for the data analysis

Pilot	Type	Protocol	Frequency
UCL (UK)	University	MQTT	On demand
EDF (France)	Offices	Nemocloud / Ellonasoft / Wattsense APIs	10 min
IASI (Romania)	Various types	InfluxDB	10 min
VEOLIA (Spain)	Residential	FTP	15 min
EMOT (Italy)	Offices	API	On demand
FOCCHI (Italy)	Factory	SolarEdge / JotMotiqa APIs MQTT Google Drive API	15 min
HERON (Greece)	Various types	API	5 min
FVH (Finland)	Various types	HTTP request	10 min
IEECP (The Netherlands)	Schools	MQTT Netatmo API	On demand
NTUA (Greece)	University	PostgreSQL	5 min

Data quality and correctness methodological approach

Two primary approaches are considered for data quality assurance: one focuses on assessing dataset quality, while the other involves data correction techniques such as interpolation and cleansing. While this work is focused on the correction techniques, the data quality evaluation is summarized from a previous manuscript (Hernández, 2024), named data quality checker.

The data quality checker operates through two distinct methods, as illustrated in Figure 2. The first method involves processing manually uploaded datasets, such as CSV files, Excel sheets, or JSON documents from pilot buildings. A dedicated data ingest ion module processes these raw files, converting them into a standardized format through a filtering mechanism. This transformation aligns the data into a structured format (e.g., a Python DataFrame), ensuring uniform representation. Additionally, users can provide a configuration file specifying parameters and relevant details for processing.

The second method leverages automated data collection technologies, such as Pentaho (Hitachi, 2023) and scripts written in Python or Java. These tools include built-in data ingestion and filtering capabilities, allowing direct integration with the subsequent analysis module.

Regardless of the input method, all processed data is passed through a dedicated statistics module, which evaluates data quality across various dimensions. This module is designed for interoperability and scalability, accepting standardized DataFrames to ensure consistent quality assessment across different data sources. By maintaining a structured and replicable approach, this methodology enables reliable evaluation of data quality across diverse datasets, ensuring consistency in analysis and reporting.

Baseline for data quality

Data quality is a key aspect and maturity levels are impacting in the final results, such as the analysis realized by (Hernández, 2022) in the cities of Nantes, Hamburg and Helsinki. One of the applications is building energy performance, under the uncertainty of data quality (Morewood, 2023).

Table 3: Data quality baseline values

Pilot	Completeness	Consistency	Accuracy
UCL (UK)	61.45 %	76.05 %	98.89 %
EDF (France)	75.70 %	63.26 %	93.47 %
IASI (Romania)	98.17 %	69.56 %	95.50 %
VEOLIA (Spain)	94.57 %	64.41 %	97.10 %

EMOT (Italy)	61.74 %	46.77 %	86.10 %
FOCCHI (Italy)	56.11 %	42.56 %	93.26%
HERON (Greece)	66.47 %	100 %	58.73 %
FVH (Finland)	85.03 %	66.26 %	91.39 %
IEECP (The Netherlands)	88.12 %	100 %	73.02 %
NTUA (Greece)	53.67 %	50.05 %	76.84 %

Under this section, a summary of the data quality baseline is made in the pilot buildings described before. Table 3 includes the completeness, accuracy and consistency dimensions. Reliability is the cartesian product between completeness and accuracy; therefore, it is not providing any added value, whereas relevance, accessibility and timeliness are out of the scope of this manuscript.

As observed, data quality differs pilot from pilot. Completeness dimension is very scattered, but reaching very high values such as IASI, while NTUA is providing low-quality values. Consistency is usually high, meaning normal distribution of data. Finally, accuracy is demonstrating many outliers are obtained from field.

Data corrector

Machine-learning techniques demonstrate being effective and efficient in the building sector, as applied by (Seyedzadeh, 2020). According to this advantage, the data corrector is designed to enhance key data quality metrics—specifically consistency, accuracy, and completeness—by leveraging machine learning techniques. The process begins with the removal of duplicate data entries, followed by an assessment of data completeness. If completeness falls below 90%, the correction mechanism is triggered to fill in the missing information.

Operating asynchronously in the background, the service activates only when a sufficient volume of data is accumulated, ensuring minimal impact on system performance and user experience. The selection of the minimum data requirements depends on performance metrics, including handling class imbalance, varying prevalence, or asymmetric cost–benefit trade-offs (Varoquaux et al., 2023). In the context of building management, activation is based on operational cycles to ensure representative conditions. For instance, space heating data requires a full winter cycle to capture stable operational patterns, whereas cooling systems may need summer data to achieve meaningful corrections. For other applications like charging stations, typically, at least 30 days of operational data are required to ensure model stability.

The architecture for the data corrector is depicted in Figure 3. It is composed of multiple components. Firstly, the configuration module allows users to specify variables for correction. The learning module retrieves historical data from the Data Warehouse (DWH) to refine its predictive models.

In order to correct data, this approach leverages different machine learning models depending on the nature of the variables. On the one hand, a linear regression model is applied when cumulative variables are aimed at being interpolated. This is the case of energy consumption, whose values are continuously increasing. The results of this model ensure consistency in progressive measurements. On the other hand, a random forest regression model is used for non-cumulative variables, such as temperature. This algorithm handles non-linearity to capture complex relationships. Both models were chosen based on their ability to balance accuracy and computational efficiency while ensuring reliable data correction.

The application of these algorithms populates new datasets to complete the original registers. Once completed, data quality indicators can be recalculated to quantify improvements and validate the correction process.

This structured approach ensures that datasets remain accurate, consistent, reliable, complete, and optimized for downstream applications. To maintain dataset integrity, newly generated data is labeled as "generated data," clearly distinguishing it from original records.

The duplicates filter processes and formats the data into a structured dataframe before passing it to the KPI Calculator module. Alongside the dataframe, sensor metadata, such as sampling frequency and timestamp range, is provided to assess completeness. The expected number of data intervals is calculated and compared to the actual records, with completeness determined as the ratio between the two.

If completeness drops below 90%, the data corrector module applies machine learning techniques to fill gaps, as explained in the next section. Once corrected, the data is evaluated for accuracy and consistency.

The Data Completeness module follows a set of key steps. It determines the appropriate model based on the data corrector configuration: a linear regression model for cumulative variables, a random forest classification model for categorical variables, or a machine-learning random forest regression model for other non-cumulative variables. It retrieves historical sensor data and identifies timestamps for missing values since the last execution.

Due to limited historical data, the dataset is split into 90% of data samples for training and validation, while 10% of data are used for testing. Model performance is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and R^2 values, with an R^2 threshold of 0.7 required for validity)a target achieved by most trained models).

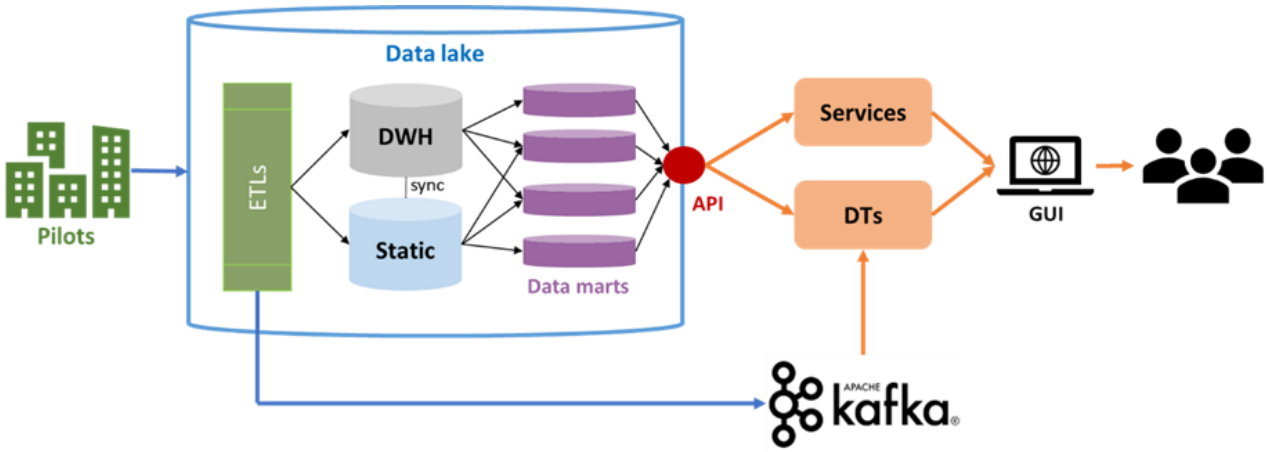

Figure 1: Data management procedure from data collection to data usage

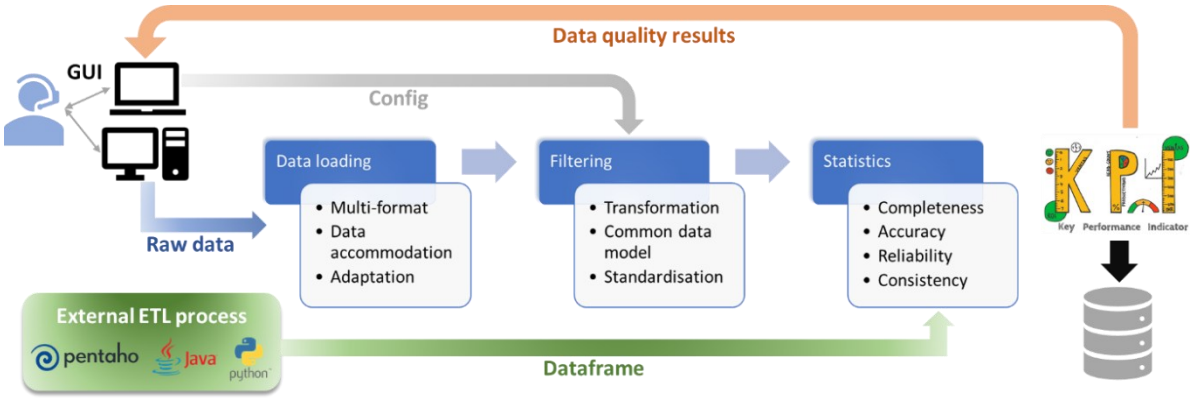

Figure 2: Data quality calculation procedure

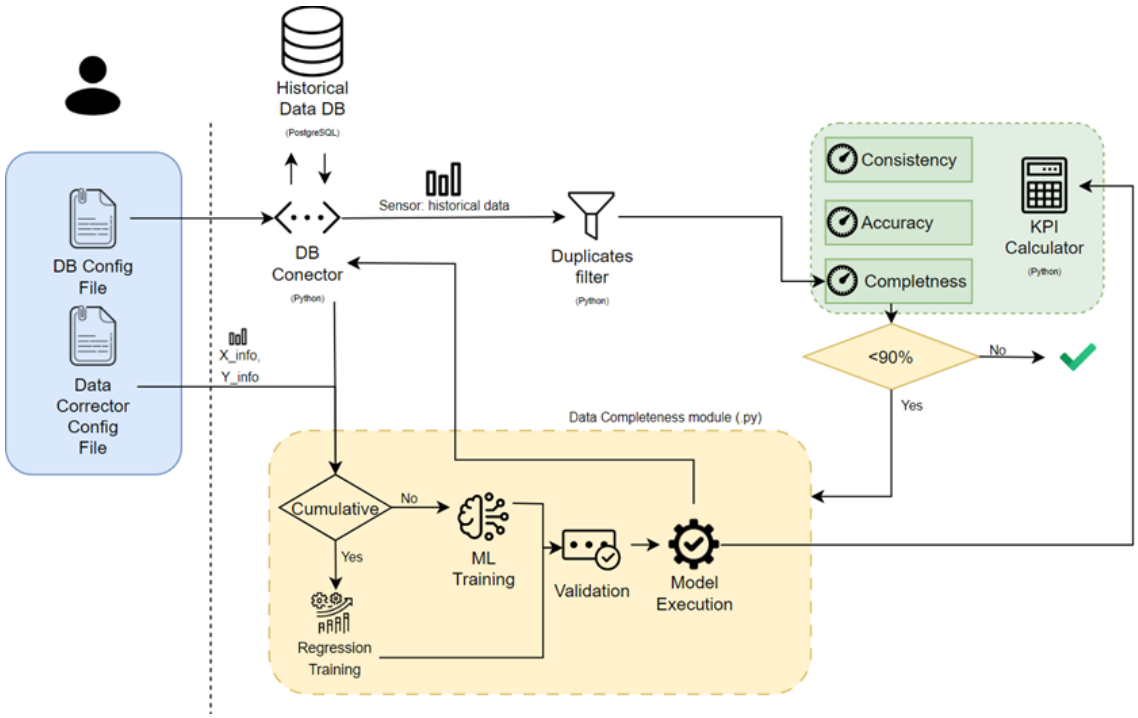

Figure 3: Data corrector components diagram

Data correction results and discussion

The models for validation of the approach have been executed for the "SENSOR_OUTDOOR_temperature" dataset from the NTUA Pilot. Figure 4 shows the original distribution of the sensor data. As observed, there exist several gaps in the data distribution along time.

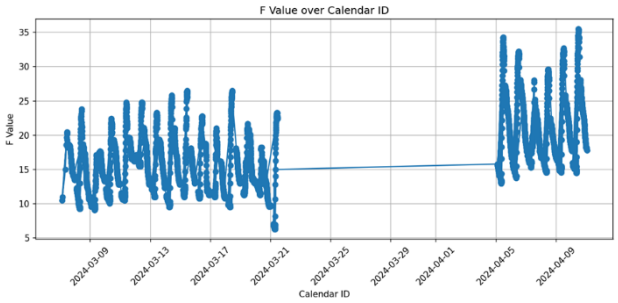

Figure 4: Dataset with gaps in the period between 09-03-2024 and 09-04-2024

Figure 5 draws the final distribution, where original values are shown in blue and predicted are shown in red. The model is then predicting "past" values (i.e., the objective is interpolating missing data) according to the trained models.

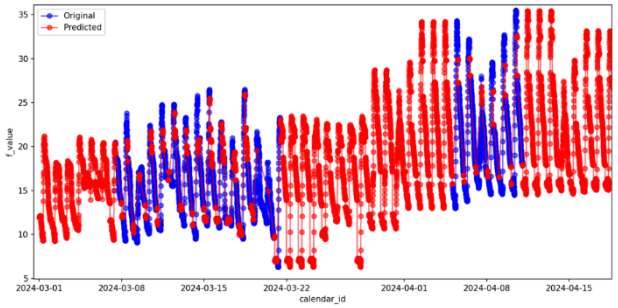

Figure 5: Corrected dataset in the period between 09-03-2024 and 09-04-2024

One of the challenges in evaluating the model's performance is the absence of ground truth data for missing values. Since these values were not recorded, direct comparison with real measurements is not feasible. Instead, a validation approach was adopted using historical data. To assess the reliability of the predictions, a portion of the available data was removed (30% of the database recorded values) and later reconstructed using the machine learning model (obviously trained without that missing data). The predicted values were then compared with the original measurements using standard error metrics, getting MSE=0.049, MAE=0.112 and R^2=0.997, demonstrating the trained model accurately reconstructs missing values with limited error.

The visual analysis indicates that consistency and accuracy metrics are ensured. The most frequent value is around 16ºC, with most samples falling within the confidence range (i.e., mean ± standard deviation). Meanwhile, in terms of accuracy, all samples remain within the defined accuracy interval, which is determined by the expected minimum and maximum values.

Quantitatively, the results confirm significant data quality improvements. The completeness metric increased from 53.67% to 100% after applying the model. Similarly, accuracy and consistency saw improvements, reaching 72.14% and 100%, respectively.

While the dataset initially presented a 53.67% completeness rate, meaning that nearly half of the measurements were missing or unreliable. According to Table 3, it highlights the same situation of multiple buildings across Europe, falling into low data quality levels. This implies uncertainty in the decision-making processes. For instance, facility managers had to rely on incomplete datasets, limiting the energy consumption analysis with potential overestimation or underestimation of HVAC loads. This lack of reliable data could result in inefficient energy use, increasing costs and reducing overall system performance.

In practical applications, unreliable data forces organizations to incorporate larger safety margins in their operations, which often translates to increased energy costs and suboptimal scheduling of maintenance tasks. By contrast, high-quality datasets enable data-driven decision-making, reducing operational risks and improving compliance with energy efficiency regulations such as the EPBD. This drives to financial savings with more precise energy use forecasting reduces unnecessary energy consumption and regulatory compliance.

Conclusions

This study emphasizes the need to improve data quality in the smart building sector. The current challenges for a better-informed decision-making process make this work highly applicable. A two-step approach is made: data quality checker and quality corrector. The methodology led to notable improvements in completeness, accuracy and consistency. These gains show how structured data management can transform unreliable information into more valuable insights.

The approach is deployed by using a centralized data lake to keep record of both dynamic and static data set, assuring data synchronization between both repositories. This setup enables interoperability and integration with digital twins and other building-related services.

Improving data quality not only boosts operational efficiency but also reduces risks in data governance. The process supports better decision-making and promotes clearer communication among stakeholders. In short, data quality improvements leads to tangible benefits, bridging the gap between digital transformation and real-world energy management outcomes.

While the methodology has proven effective, future work should test its performance in real-time applications, particularly for dynamic environments. Tailoring the correction strategies to specific building categories could further improve outcomes. Expanding the framework to include factors like timeliness and accessibility would also strengthen its usefulness.

Acknowledgments

This work has been made under the DigiBUILD EU project, funded under the GA#101069658, and Auto-DAN EU project under GA# 101000169 of the Horizon Europe programme.

During the preparation of this work, the authors used DeepL and ChatGPT to improve the grammar and readability. After using these tools, text was reviewed and edited. Authors take full responsibility for the content of the publication.

References

Abraham, R., Schneider, J., Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 49, pp. 424-438.

Alsuhaibani, A., Han, B., & Leite, F. (2022). Investigating the causes of missing field detected issues from BIM-based construction coordination through semistructured interviews. Journal of Architectural Engineering, 28(4), 04022028.

BDVA/DAIRO position paper. (2021). Response to the European Commission's proposal for AI Regulation. https://www.bdva.eu/sites/default/files/BDVA_DAIRO%20response-feedback%20AI%20Regulation_Final.pdf, visited on 10th January 2024.

Chopade, A.S., et al. (2024). Enhancing Sustainability in Energy Efficiency Programs: Leveraging Data-Driven Mitigation Strategies and DLT to Address ESCO Challenges. Proceedings of Third International Symposium on Sustainable Energy and Technological Advancements. ISSETA 2024. Lecture Notes in Electrical Engineering, vol 1254. Springer, Singapore. https://doi.org/10.1007/978-981-97-7018-2_10.

Ender, L. 'Data Governance in Digital Platforms: A case analysis in the building sector', Dissertation, 2021

EPBD (2023) https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en#current-rules-to-improve-the-eus-building-stock

Hernández, J. L., García, R., Schonowski, J., Atlan, D., Chanson, G., & Ruohomäki, T. (2020). Interoperable Open Specifications Framework for the Implementation of Standardized Urban Platforms. Sensors, 20(8), 2402.

Hernández, J.; Quijano, A.; García, R.; Nouaille, P.; Risch, L.; Virtanen, M. and de Miguel, I. (2022). Analysis of Data Quality in Digital Smart Cities: The Cases of Nantes, Hamburg and Helsinki. In Proceedings of the 11th International Conference on Data Science, Technology and Applications - DATA; ISBN 978-989-758-583-8; ISSN 2184-285X, SciTePress, pages 353-360.

Hernández, J., Martín, S., Kapsalis, P., Katsigarakis, K., Sarmas, E., Marinakis, V. Building a Data Lake for Smart Building Data: Architecture for data quality and interoperability, 2023 14th International Conference on Information, Intelligence, Systems and Applications (2023a), University of Thessaly, Greece, 10-July 2023.

Hernández, J. L., Martín, S., Marinakis, V., & de Miguel, I. (2023b). From silos to open, federated and enriched Data Lakes for smart building data management. In 2023 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv) (pp. 29-33). IEEE.

Hernández, J. L., Martín, S., Lonac, V., de Miguel, I. & Rossi, A. (2024). Data Quality in The Built Environment Across Europe. Proceedings of the 2024 European Conference on Computing in Construction. DOI: 10.35490/EC3.2024.217.

Hernández, J.L., de Miguel, I., Vélez, F., Vasallo, A. (2024b). Challenges and opportunities in European smart buildings energy management: A critical review. Renewable and Sustainable Energy Reviews, Volume 199, 114472, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2024.114472.

Hitachi, Pentaho Data Integration, online: https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.1/Products/Pentaho_Data_Integration, visited on 27th December 2023.

Hossein Motlagh, N.; Mohammadrezaei, M.; Hunt, J.; Zakeri, B. Internet of Things (IoT) and the Energy Sector. Energies 2020, 13, 494. https://doi.org/10.3390/en13020494.

Koo, H. J., (2024). Clash relevance prediction in BIM model coordination using artificial neural network. Construction Research Congress 2024, 127-136.

Morewood, J. (2023). Building energy performance monitoring through the lens of data quality: A review, Energy and Buildings, Volume 279, 2023, 112701, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2022.112701.

Seyedzadeh, Saleh, Rahimian, Farzad Pour, Oliver, Stephen, Rodriguez, Sergio and Glesk, Ivan, Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making, Applied Energy, Volume 279, 2020, 115908, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2020.115908.

Varoquaux, G., Colliot, O. (2023). Evaluating Machine Learning Models and Their Diagnostic Value. In: Colliot, O. (eds) Machine Learning for Brain Disorders. Neuromethods, vol 197. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-3195-9_20

Yong Teng, S., Touš, M., Dong Leong, W., Shen How, B., Loong Lam, H., Máša, V. Recent advances on industrial data-driven energy savings: Digital twins and infrastructures, Renewable and Sustainable Energy Reviews, Volume 135, 2021, 110208, https://doi.org/10.1016/j.rser.2020.110208.