Smart and Sustainable Built Environment

Towards environmentally enriched floor layout datasets: a workflow for transitioning the existing data in the built environment

Fatemeh Mostafavi, Seyran Khademi and Georg Vrachliotis
Delft University of Technology, Delft, Netherlands

Received 31 October 2025
Revised 16 February 2026
Accepted 14 March 2026

Abstract

Purpose – This paper aims to present data refinement and enrichment workflow to integrate building performance guidelines with existing semi-structured floor layout datasets. The goal is leveraging the application of architectural datasets in the built environment across data-driven methods as well as enabling informative visualizations and large-scale analyses.

Design/methodology/approach – The Swiss dwellings dataset is employed as the foundation in this study, which later undergoes a Python-based data refinement, feature engineering and attribute extension. The modified attributes cover spatial zoning (categorical), proxy indicators for daylight metrics and view layers (numerical), noise level (numerical), acoustic comfort (categorical) and window orientations (categorical).

Findings – The study presents an efficient workflow of turning textual data of the building performance guidelines into structured tabular data suitable for machine learning. Moreover, the visualizations of the structured floor layouts data reveal new insights as a result of analyzing the dataset. The Oriented Environmental Swiss Dwellings (O-ESD) dataset, as the main product of this study, brings data-driven learning opportunities from existing floor layout datasets towards environmental design automation. Moreover, O-ESD offers human-interpretability through the structured micro-climatic visualizations.

Originality/value – There has been no previous effort in the field for upgrading the existing architectural datasets in alignment with the building performance guidelines to expand their applicability in data-driven approaches. The proposed workflow not only gives insights into data refinement applications in the field but also results in an environmentally enriched floor layout dataset as the outcome. The resulting dataset, the workflow towards it and example visualizations are released publicly.

Keywords Floor layout datasets, Machine learning, Data refinement, Building performance guidelines, Environmental building design, BREEAM

Paper type Research article

1. Introduction

At each phase of architectural design, from pre-design and schematic design to design development and construction administration, different tasks and requirements are addressed (Waters *et al.*, 2023). At the early design stages, the cost of decision changes is lower and the ability to impact functional capabilities is higher (MacLeamy, 2004). To make informed decisions and effectively transform traditional design principles into more intelligent design processes, access to trustworthy data is crucial (Lystbæk, 2025). Data collection, labeling, cleaning and analyzing have shown heightening downstream impacts in data-driven workflows (Sambasivan *et al.*, 2021). Access to rich, complete and homogeneous data has been and still is a central challenge, particularly when data are sourced from historical records, experiments and simulations (Reich, 1997). Therefore, to be able to effectively employ data in different phases of architectural design by both designers and machines, raw data should undergo pre-processing and special treatment including removing redundant or incorrect data,

Smart and Sustainable Built Environment
Emerald Publishing Limited
e-ISSN: 2046-6102
p-ISSN: 2046-6099
DOI 10.1108/SASBE-10-2025-0684

that is data refinement. Moreover, extensive and informative data about data, that is metadata further enriches the associations among data and enhances the interpretation of the design decisions (Stouffs *et al.*, 2004).

Design automation in architecture has been historically perceived as a means of freedom if being dealt creatively and sensibly (Vrachliotis, 2022). Recent advancements in design automation increasingly rely on machine learning (ML) algorithms to enable data-driven design tools. Data are being leveraged by researchers across nearly all domains for ML and other AI applications (Greenberg *et al.*, 2025); however, lack of well-distributed and contextually rich data, that is metadata may result in large errors in the learning-based models. In the context of design automation, the large-scale datasets have not been curated purposefully for architectural qualities and have been utilized merely based on the availability (Weber *et al.*, 2022). This gap needs to be bridged by developing brand new datasets with relevant attributes and enriching the existing ones through attribute transformation and addition. The metadata features could consequently act as supervision signals used to train AI-based models (Sucholutsky *et al.*, 2023). Examples of metadata in the schematic design phase, when design optioning occurs and major environmental decisions are made, include daylighting, view and noise conditions within the building's context. The outdoor micro-climatic attributes have also shown significant impacts on the indoor thermal and daylight comfort of occupants, specifically in medium-density housing (Pourel *et al.*, 2025).

Most of the commonly used floor layout datasets for ML applications, such as RPLAN (Wu *et al.*, 2019) and LIFULL HOME's (LIFULL Co. Ltd, 2015) mainly contain mere geometrical attributes; hence, lack the environmental measures, limiting the capacity to develop ML models with environmental awareness. Given the availability of independent environmental guidelines for buildings, integration of environmental metadata with floor layout data is plausible. Including environmental-based design and evaluation guidelines, developed for assuring the occupants' health and wellbeing in the built environment has not been a straightforward task. Therefore, studies have been conducted to integrate the measurable attributes embedded in the guidelines to reach real-time assessment on the building data, which has been commonly represented in building information models (BIM) (Chelaru *et al.*, 2025; Kamsu-Foguem *et al.*, 2019). The methodological gap here is to efficiently read and transform building specific environmental attributes into large-scale building datasets.

Big data, in its most refined form, is typically organized into structured tables or grids, such as comma-separated values (CSV) files (van den Burg *et al.*, 2019). In our approach, the floorplan data for guideline compliance is prepared by integrating the geometrical and environmental attributes of residential buildings into structured data frames aligned with the selected building performance criteria. In this study, multiple geometrical and environmental attributes of the Swiss Dwellings floor layout dataset (Standfest *et al.*, 2023) are transformed to new attributes including spatial zonings, orientations, proxy illuminance factors and view layers, noise levels and acoustic comfort. These attributes are mainly derived from the typology- and climate-aligned (with respect to the Swiss Dwellings) building performance guidelines including BREEAM International, New Construction (BRE Global, 2021), EN 17037:2018 + A1:2022 (European Committee for Standardization, 2021) and Night noise for Europe (WHO, 2009). In the process of data refinement and enrichment, the proposed workflow in this study addresses the following questions.

(1) How did the advancements in computational technologies and data-driven methods influence the process of building regulation compliance checking?

(2) What is the approach for step-by-step data refinement and attribute transformation that could lead to an environmentally enriched dataset with enhanced usability?

(3) What metadata are required for architectural datasets to accommodate for AI-assisted environmental generation and assessment of residential floor layouts?

By introducing a workflow that translates building performance guidelines into attributes of large-scale building datasets, this study contributes to the creation of an environmentally enriched floor plan dataset, supporting the development of environmentally aware intelligent design tools.

The rest of the manuscript investigates the related work in the field of floor layout data, building performance guidelines and guidelines compliance checking workflows in section 2. Consequently, section 3 explains the proposed workflow of enriching the existing floor layout datasets guided by building performance guidelines. Next, section 4 sheds light on the findings of this workflow and reports the results of the study through analyzing the enriched dataset, O-ESD and its attributes, limitations and possible downstream applications. Ultimately, chapter 5 concludes the study by summarizing the necessity, method and results of the conducted research.

2. Related work

2.1 Floor layout data

Architectural floor layout data plays a critical role within a bigger framework of design and evaluation in an interplay with the machine and the designer (Mostafavi *et al.*, 2025). After the first paradigm shift from conventional design to rule-based design, the second shift towards AI-based architectural design clearly called for big data (Zhuang *et al.*, 2025). However, besides diversity in style and lack of annotation consistency (Pizarro *et al.*, 2022), absence of floor layout environmental attributes has made the usability of the floor layout datasets challenging (Mostafavi and Khademi, 2023). RPLAN (Wu *et al.*, 2019), as the most used dataset in the related studies, only contain geometrical attributes of simple-shaped and axis-aligned floor layouts. More specifically, RPLAN dataset is released in the form of four-channel bitmap images, the first one representing exterior walls and the front door, the second one representing room types, internal walls and doors, the third one representing space indices and the fourth one outlining the floor layout's boundary. LIFULL HOME's dataset (LIFULL Co. Ltd, 2015), as the second most used dataset in the related studies contains rental metadata (rental fee, area, location, age, structure and facilities) and image data (floorplan image, room view, etc.) in the tab-separated values (.tsv) and joint photographic experts group (.jpeg) file format, respectively. As a consequence of data availability condition, the problem formulation, intended task and evaluation criteria of the studies based upon these datasets were inevitably bound to what the dataset attributes could afford (Hosseini *et al.*, 2023; Hu *et al.*, 2020; Luo and Huang, 2022; Nauata *et al.*, 2020, 2021; Shabani *et al.*, 2023; Shim *et al.*, 2024; Sun *et al.*, 2022; Tang *et al.*, 2023; Upadhyay *et al.*, 2022). In other words, due to the lack of metadata on spatial zoning, daylighting and view conditions, noise levels and orientation of windows, there have been no previous studies using the RPLAN dataset in which the building environmental condition was included.

The Swiss Dwellings dataset (Standfest *et al.*, 2023) is currently the only available large-scale floor layout dataset featuring both geometrical and environmental attributes. The dataset includes data of over 45,000 apartments in about 3,100 distinct buildings in the climatic context of Switzerland. Data of Swiss Dwellings are released in four CSV files named "geometries", "location ratings", "locations" and "simulations". There have been previous studies exploring the Swiss Dwellings dataset across various applications including feature selection (Bielik *et al.*, 2023), micro-climatic building context visualization (Mostafavi and Khademi, 2023), multimodal learning in visual comfort assessment (Kat *et al.*, 2024), wayfinding quality assessment (Alsaggaf, 2024) and modification of the dataset for the task of floor plan autocompletion (van Engelenburg *et al.*, 2025).

A preliminary step of data refinement was included in some of the above-mentioned studies to different degrees depending on the downstream task. In the study conducted by (Mostafavi and Khademi, 2023) the main goal of data refinement was merging the geometrical and environmental data of the Swiss Dwellings dataset, whereas a more extensive modification with the exclusion of simple layouts was performed in the study by (van Engelenburg *et al.*, 2025), which led to the release of the Modified Swiss Dwellings (MSD) dataset. Moreover, using the attributes both directly from the Swiss Dwellings dataset and indirectly through the isovist assessments (Alsaggaf, 2024), curated a version of the original dataset with attributes of 256 living rooms for the wayfinding quality evaluation. Although the mentioned studies explored certain aspects of this extensive dataset, the potentials of environmental simulation data of Swiss Dwellings have not been thoroughly investigated yet due to its novelty, complexity and cross-document misalignments. Accordingly, this study aims to refine, distill and merge the existing data in the Swiss Dwellings to accelerate its usability, specifically in environmental floor layout generation and evaluation tasks.

2.2 Building performance guidelines

To guide the design and assessment of buildings across different health and wellbeing measures, several guidelines, standards and certification systems have been developed. As the first sustainability assessment certification and building rating system, Building Research Establishment Environmental Assessment Method (BREEAM) was set up in 1990. Among other schemes, BREEAM curated specific guidelines for housing typology, ranging from New Homes, EcoHomes and Multi-Homes, to the more recent country-specific ones (Graaf, 2023). BREEAM New Homes was specifically curated for assessing designs of new single-household dwellings and it took the daylight provision (using the Daylight Factor metric) and view to sky of the habitable spaces into consideration (Raw and Prior, 1991). As an upgrade to New Homes, BREEAM EcoHomes brought an extensive guidance for providing acoustic comfort and avoiding noise complaints in households, as well as introducing a daylight calculation tool (Building Research Establishment Ltd, 2006). Further, BREEAM New Construction brought the complete range of building types, including the requirements of multi-residential buildings into a single scheme (BRE Global, 2016). The latest available version of BREEAM International New Construction V6.0 also encompasses credits for different aspects of health and wellbeing for various building types, including single and multiple residential dwellings (BRE Global, 2021). Besides health and wellbeing measures, BREEAM also emphasizes the importance of spatial zoning and orientation in line with passive design strategies, reference building definitions and building control design.

In the context of Europe, EN 17037:2018 + A1:2022 provides guidelines and requirements for maximizing the natural daylight and view and minimizing the glare in the buildings (European Committee for Standardization, 2021). As a determining factor in occupants' visual comfort in buildings, the quality of view towards outside is widely assessed in the related building standards concerning health and well-being (Abdelrahman *et al.*, 2023). Besides health benefits, providing high-quality view condition has shown non-visual and architectural effects on the way a user experiences a space (Hraška and Čurpek, 2024). In addition to the visual comfort, there have been guidelines to maintain the health and wellbeing of occupants according to the acoustic measure, as night noise level has been shown to have a critical effect on the night sleep quality of occupants in residential buildings (Clark and Paunovic, 2018). As one of the main references for assessing the acoustic comfort in housings in the context of Europe, the Night Noise Guidelines for Europe has been established (WHO, 2009). Despite the emphasis of health and well-being-related guidelines on visual and acoustic comfort as well as spatial zoning and orientation measures, there are no available floor layout datasets which provide numerical or categorical attributes useful for large-scale ML-based design or assessment purposes.

2.3 Guideline compliance checking workflows

There has been an active effort in the previous studies regarding automating the process of regulations, guidelines or building codes compliance checking in the built environment. Multiple studies investigated ontology-based workflows in which attributes of the BIM models and intended regulations are mapped to enable comparison (Amor and Dimyadi, 2021; Hagedorn *et al.*, 2023; Kamsu-Foguem *et al.*, 2019; Zhong *et al.*, 2018). More specifically concerning building sustainability aspects Kamsu-Foguem *et al.* (2019), proposed a compliance checking approach using mapping of conceptual graphs for a selection of BREEAM requirements and the target BIM based on ontology taxonomies. In a broader workflow suggested by Chelaru *et al.* (2025), BIM acting as the building data source, BREEAM, Autodesk Forma (Autodesk Forma, 2025) and Dalux CDE (Dalux, 2025) were integrated to check the alignment of an educational building design parameters with BREEAM criteria in health and wellbeing category (daylight potential, sun hours and noise), energy and land and ecology. Authors emphasized ensuring the presence of accurate information in the BIM model in compliance with the selected BREEAM category requirements, which resulted in an extensive data collection and pre-processing step. More recently, researchers explored the potential of MSD dataset in a joint floorplan generation and the accessibility code compliance check (Zhang and Zhang, 2025).

With the evolution of computational technologies and the surge of data-driven approaches, the workflow of integrating building regulations in data processing requires an upgrade to accommodate for current data formats in the large-scale existing datasets. As the digital representation of the design data, BIM has been repeatedly included in the regulatory compliance checking workflows in the previous body of research. However, many building design specifications required for BIM representation of building data are not definite at the early design stages, and BIM data are too complex to be seamlessly adopted in compliance checking workflows (Beach *et al.*, 2020). The link of BIM models with textual data has been previously explored in the scope of defect reports, project communication and product model (Sobhkhiz and El-Diraby, 2023). However, analyzing the environmental performance of buildings has not been the central focus of the related studies, since the building performance measures are not stored in BIM or IFC format.

Applying environmental data-driven methods on BIM data is highly constrained by limited number of buildings due to the complexity and comprehensiveness of data, as well as being reliant on external software or sensors (Kanna *et al.*, 2022; Shen and Pan, 2023). Also, previous research has shown that noise in BIM data instances hinder the semantic information alignment checking (Zhou and El-Gohary, 2021). In this sense, data models need migration to data formats such as data frames which are operable in data-driven approaches. Plus, the integration of BIM data with building performance standards has resulted in a complicated system of various tools such as Revit, Forma, Dynamo, AutoCAD, DesignBuilder, etc., which are not all open-source and free to use.

To our best knowledge, there are no previous efforts regarding the integration of sustainability certification's guidelines with existing large-scale datasets in architecture. Moreover, the lack of a large-scale dataset curated appropriately for addressing environmental impact has been repeatedly expressed (Mostafavi and Khademi, 2023; Weber *et al.*, 2022). Therefore, the current study aims to enrich one of the existing building datasets, Swiss Dwellings, using the attributes needed to design and evaluate residential floor layouts guided by three sources of building performance guidelines.

3. Methods and materials

Situated at the schematic design phase, where major environmental design decisions are to be made, this research uses an existing dataset as a foundation for applying the proposed data refinement and enrichment workflow (Figure 1). It starts with two semi-structured tabular data sourced from the Swiss Dwellings dataset as two separate CSV files (geometries and simulations). Then, three building performance guideline sources (BREEAM International New Construction, EN 17037, and Night noise guideline for Europe) in the structured textual data form are used as a guide for data refinement, attribute transformation and attribute addition applied on the Swiss Dwellings dataset. The guidelines sources are selected based on

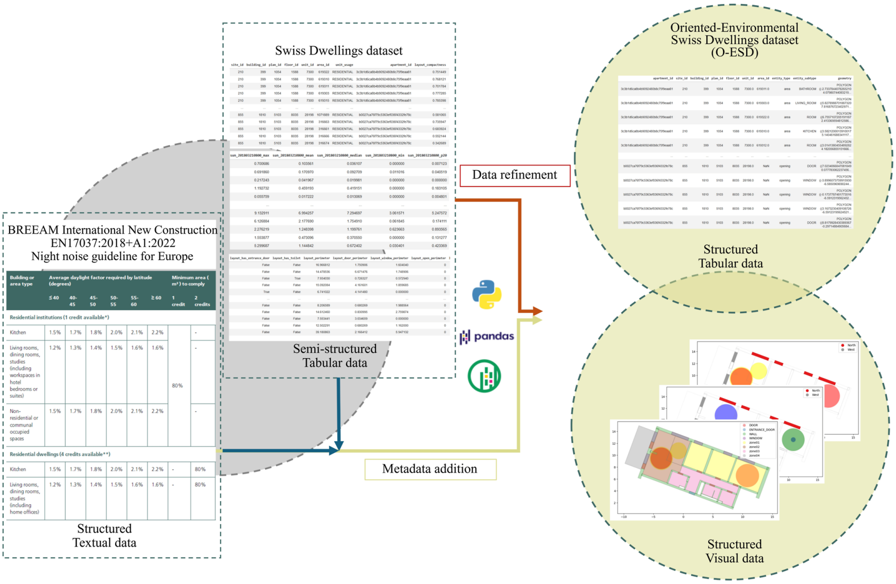

Figure 1. The proposed workflow for transitioning existing datasets towards environmentally enriched data in the built environment, here applied on the Swiss Dwellings dataset

the compatibility with the Swiss Dwellings' layout typology (i.e. residential) and its climatic context (i.e. Europe). Multiple Python libraries including Pandas and GeoPandas are employed to make the transition towards the environmentally enriched dataset, O-ESD, which is a structured tabular data in the format of a data frame. Additionally, the workflow results in structured visual data containing floor layout and window orientation maps of each apartment unit in the O-ESD dataset. The following sub-sections walk through the data enrichment steps and the dataset's environmental impact.

3.1 Dataset enrichment

In this study, the enrichment of the Swiss Dwellings dataset is divided into three main stages: (1) data refinement, (2) attribute transformation and (3) attribute addition. The following paragraphs will walk through the specifications of these stages. The guideline-inspired requirements which could be inferred from the existing raw data in Swiss Dwellings are brought in summary in Table 1. The requirements regarding the daylight provision and view quality are presented under the same category of "visual comfort", and more view layers such as view to landscape and city are added to conventional view to sky requirements. The limit values are picked based on the availability of data in the Swiss Dwellings dataset.

3.1.1 Data refinement. For data to be processed neatly, a refinement procedure on the raw data frames was followed. Applying the data refinement on the Swiss Dwelling started with removing duplicate rows from both geometry and simulation data frames. Next, since this study focuses on housing typology, the layouts with any other unit usage except for residential were excluded. Consequently, the entity type of "feature" was omitted, leading to the remaining entity types of "area", "separator" (walls and columns) and "opening" (door, windows and entrance door). Next, the entity subtype of "railing" was excluded from the geometry data frame, since it only outlines the exterior perimeter of balcony area and is not central to the aim of this study. After taking these measures, the dataset was ready to undergo the attribute transformation process.

Table 1. Indicators inspired from the building performance guidelines addressed in the current study

Health and wellbeing aspect	Assessable criteria by O-ESD	Source
Visual comfort – daylight provision	Minimum average daylight factor 1.8% in kitchen and 1.4% in the other living spaces (living room, dining room, study room and home offices) for the corresponding latitude range of Switzerland at the minimum of 80% of the area	BREEAM International New Construction V6.0 (BRE Global, 2021)
Visual comfort	At least 80% of the rooms have a view to sky from a desk or table top (height 0.85 m in residential buildings)	BREEAM International New Construction V6.0 (BRE Global, 2021)
– view quality	The number of view layers visible from 75% of the evaluated spaces	EN 17037:2018 + A1:2022 (European Committee for Standardization, 2021)
Acoustic comfort	Maximum noise level (LAmax) of 32 dB in the bedrooms during nighttime hours	Night Noise Guideline for Europe (WHO, 2009)
Spatial requirements	Assessable criteria by O-ESD	Source
Spatial zoning	Building control (lighting, ventilation)	BREEAM International New Construction V6.0 (BRE Global, 2021)
Orientation	Reference building definition Passive design strategies Building energy type definition	BREEAM International New Construction V6.0 (BRE Global, 2021)

3.1.2 Attribute transformation. With an eye on the selected building performance guidelines, some raw attributes already existing in the simulation data of Swiss Dwellings dataset were transformed into the proxy attributes that could indicate some aspects of visual and acoustic comfort. The steps taken are as follows:

Grouping spaces: After completing the first round of data cleaning, the remaining entity subtypes included “bathroom”, “living room”, “room”, “kitchen”, “balcony”, “corridor”, “wall”, “door”, “window”, “entrance door”, “dining”, “shaft”, “storeroom”, “column”, “living dining”, “staircase”, “bedroom”, “outdoor void”, “loggia”, “void”, “kitchen dining”, “terrace”, “wintergarden”, “lightwell”, “elevator”, “patio”, “garden”, “technical area” and “studio”. These entity subtypes belong to entity types of “area”, “openings” and “separator” and all under the “residential” unit usage. Next, these entity subtypes were grouped together based on their environmental demands. In this sense, spaces demanding morning daylight and noise avoidance (room, bedroom and studio) were grouped into zone 1, spaces requiring afternoon daylight and sufficient view to outdoors (living room, kitchen, dining spaces) were grouped into zone 2, interior spaces with no environmental demand (corridor, storage, staircase and bathroom) were grouped into zone 3, the outdoor spaces that act as a noise and thermal barrier (balcony, loggia, terrace, garden, etc.) were grouped into zone 4 and the remaining spaces (shaft, void, lightwell, patio, elevator and technical area) were labeled as “remaining”.

Transforming daylight data: As the evaluative metrics to assess the daylighting condition in an indoor space, two proxy daylight metrics were added to the O-ESD attributes. The calculation method of these two metrics was inspired by the BREEAM International New Construction guidelines for the visual comfort assessment, in which the minimum values of average Daylight Factor required for residential dwellings should comply to at least 80% of the assessed area. Daylight Factor represents the ratio of interior illuminance at a point within a building to the exterior horizontal illuminance under an overcast sky (Sun *et al.*, 2018).

Given the data availability, some choices were made to transform the raw daylight data of the Swiss Dwellings dataset to the proxy indicators aligned with the BREEAM criteria. Firstly, the simulation results of illuminance corresponding to the day of December 21st were selected for processing as a more accurate representation of the overcast sky condition ("Satellite climatology - MeteoSwiss", 2025). Secondly, the 80th percentile (i.e. p80) values of illuminance data in Swiss Dwellings were picked to calculate the proxy daylight metrics in accordance with the minimum percentage of area to comply with the daylight thresholds. Thirdly, the maximum value of the illuminance in the whole dataset (corresponding to an illuminance level in a balcony) was used as the closest indication of the exterior horizontal illuminance. Lastly, given that there are illuminance simulation data at four timesteps on December 21st (from 10:00 to 16:00 with two-hour interval) in the Swiss Dwellings dataset, one proxy daylight metric was defined to signify the daylight condition in the morning (at 10:00) and one in the late afternoon (at 16:00). Following the mentioned reasoning, the proposed daylight metrics were named Morning Illuminance Factor (MIF) and Afternoon Illuminance Factor (AIF), which measure the ratio of the illuminance value at 10:00 and 16:00, respectively, on 21st of December in at least 80% of the positions in an area compared to the maximum illuminance value in the whole dataset. These devised metrics are therefore more precisely indicated by $MIF_{max,80\%}$ and $AIF_{max,80\%,}$ respectively.

It is worthwhile to mention that although commonly-used daylight performance and visual comfort metrics encompass Spatial Daylight Autonomy (sDA), Useful Daylight Illuminance (UDI), spatial Glare Autonomy (sGA) and Annual Solar Exposure (ASE) (Hosseini *et al.*, 2025), none of these metrics could be directly calculated from the available raw daylight data in the Swiss Dwellings dataset. After transforming the raw daylight data into two enriched attributes, the underlying attributes containing unprocessed illuminance values were dropped from the dataset.

Transforming view quality data: Aligned with the available simulation view data in the Swiss Dwellings dataset, which expresses the visible amount of buildings, greenery, water, etc. occupying the spherical field of view, the third requirement in Table 1 was feasible to be considered in the current study. Similar to the data selection in daylight assessment, the p80 view data were selected again as the proxy indicator of the required view condition in 75% of the evaluated area. Accordingly, the raw view data were categorized as follows:

(1) Urban view layer: containing view to buildings, site, railway tracks, highways, primary streets, secondary streets, tertiary streets and pedestrians.

(2) Landscape view layer: containing view to greenery, and mountain class 2, 3, 4, 5 and 6.

(3) Ground view layer: containing view to water and ground.

(4) Sky view layer: containing view to sky.

Similar to the previous step, the underlying attributes containing unprocessed view values were dropped after transforming the raw view data into four enriched attributes.

Transforming noise data: To consider the acoustic comfort of the floor layouts, the raw acoustic simulation results of the Swiss Dwellings were selected to transform. The values presenting the noise intensity are expressed in dBA (decibels) for each room according to the traffic and train noise datasets. In these simulations, adjacent buildings were considered noise-blocking elements, and the values are provided for day and night hours. In this study, the nighttime noise level attributes were selected due to the alignment with the residential buildings' occupancy patterns. To reach the combined effect of both traffic and train noise sources, the logarithmic sum was applied for each room. As a result, the two selected noise attributes were transformed into one, called "noise night", expressed in decibel. Next, another attribute called "acoustic comfort" was added according to the noise thresholds in residential spaces specified by the Night Noise Guidelines for Europe (WHO, 2009). This attribute, acting as a classification label, carries two possible values ("pass"/"not pass") depending on whether the total noise level would fall below (i.e. "pass") or above (i.e. "not pass") the threshold in each zone type. Transforming the sound level values into categorical data is a common practice in the field of predicting occupants' satisfaction with indoor environment (Croffi *et al.*, 2025). Following the same logic and using the recommended values stated in the night noise guideline, the thresholds were determined as 32 dB for zone 1 (including bedrooms), 35 dB for zones 2 and 3 (zones adjacent to bedrooms) and 45 dB for zone 4 (outside spaces such as balconies). Given that adding the acoustic comfort labels required conditioning on the zone type, this attribute was added after merging the geometry and simulation data frames.

Merging data frames: After transforming the raw environmental attributes of Swiss Dwellings into the enriched ones, a merged data frame called O-ESD was created to reconcile the geometry to the simulation data, turning the original dataset into a richer and more practical data frame. The finest shared attribute between the two data frames is the identification number of areas, that is "area_id". This column was therefore used as the connection to comprehensively bridge the geometry data of Swiss Dwellings to the transformed environmental attributes resulting from the simulation data. After merging the two data frames, duplicated identification numbers including site_id, floor_id, unit_id, apartment_id and building_id were dropped, resulting in a unified data frame with enriched geometrical and environmental attributes. This step would make conditional operations of the simulation data with respect to the geometrical data feasible.

Removing complicated layouts: After the initial cleaning of the merged data frame, complicated floor layouts were identified and removed to facilitate the use of O-ESD in the ML-based downstream tasks in terms of spatial connectivity compatibility. After several rounds of visualizations and analysis, the apartment units containing at least one room with more than 6 windows or more than 8 doors were omitted. The removal thresholds on the number of windows and doors were decided based on the total number of rooms in a layout and also the distribution of data at every window or door number. It is worthwhile to mention that these thresholds can be conveniently tweaked depending on the downstream tasks' requirements, making the O-ESD compatible with various problem statements with different degrees of spatial complexity. Two instances of the removed complicated layouts are displayed in Figure 2.

3.1.3 Attribute addition. **Adding orientation data:** Determining the location of environmental factors in a building context with respect to the interior spaces, orientation is a significant feature in contextual design (Zawidzki and Szklarski, 2020). Since windows act as the connection between the interior and the context of a floor layout, accessing their orientation data can facilitate the relevant ML-bases environmental generation and evaluation downstream tasks. Therefore, a new attribute called "window orientation" was added to the merged data frame by taking the following steps for each unit in the O-ESD dataset having at least one area with at least one window:

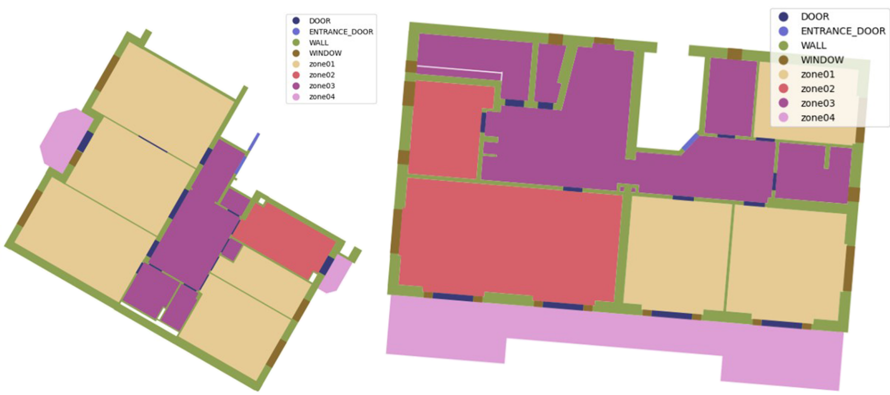

Figure 2. Examples of complicated floor plan instances in Swiss Dwellings regarding the door number (left) and window count (right), which were omitted from O-ESD

(1) The geometrical characteristics of rectangular-shaped windows were used the first indication of their orientation. Accordingly, the elongation direction of the windows was specified through comparing their length on x- and *y*-axis. As a result of this step, every window would fall into the "vertical" or "horizontal" category.

(2) The relative position of window centroids was compared to the centroid of the floor layout to specify the four main orientations. For instance, if the x value of a vertical window's centroid is larger than that of the whole floor layout, the window's orientation is regarded as East. The same logic was applied to check the West, North and South orientations.

(3) More granularity on the East and West directions was applied to achieve four new refined orientations, namely, North-East, South-East, North-West and South-West. For this purpose, the normal angle of each window was compared to the *x*-axis. Accordingly, windows with the normal angle of 30–60° fell into the North-East category, and with 120–150° into North-West category. The minus values of the same angle ranged resulted in the South-East, and South-West, respectively.

Illustrative examples of sample floor layouts and their corresponding window orientation map are demonstrated in Figure 3.

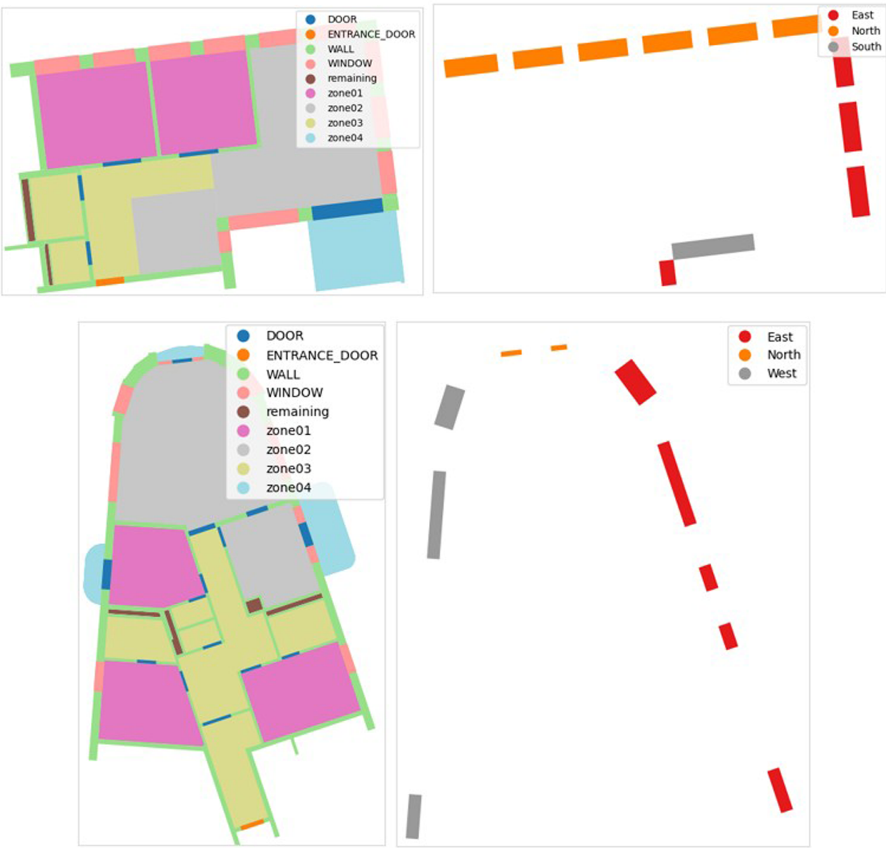

Figure 3. Window orientation demonstration for sample O-ESD floor layouts

3.2 Dataset's environmental impact

One of the main aspects of practicalities regarding working with big data is the size of the dataset and its corresponding memory usage. Lower memory equals faster execution of different data processing tasks on limited systems. Therefore, tracking memory usage helps in optimizing performance, preventing memory leaks and ensuring scalability for large datasets, which affects the energy draw and consequently, the carbon footprint of running computational operations on data (Lannelongue *et al.*, 2020). Assessing and minimizing the impact of AI -related systems on environmental sustainability and energy-efficient training of AI models through managing computational resources is also encouraged by the EU AI Act (Office of the European Union, 2024).

In the process of refinement and enrichment of the Swiss Dwellings dataset in this study, CPU memory usage (RAM) was utilized as the hardware accelerator. To clarify the impact of each step along this process, a percentage decrease for the number of rows, columns and accordingly, memory usage is reported in Table 2 in a step-by-step manner. Comparing the combined effect of geometry and simulation data frames of Swiss Dwellings with the final O-ESD dataset shows 49.54% reduction in memory usage (Figure 4). This would facilitate further downstream tasks in terms of execution speed and wider applicability on various system configurations with less energy draw.

4. Findings and discussion

4.1 O-ESD analysis

O-ESD, enriched with the attributes aligned with the selected requirements in BREEAM International New Construction, EN 17037, and Night noise guideline for Europe showcases a compact yet comprehensive floor layout dataset. This dataset provides 31 identifying, spatial and environmental attributes of more than 240,000 rooms in 44,000 apartment units of 3,000 buildings located in more than 1,400 unique sites. The units vary in floor number in the range of −4 to 20, representing the diversity in terms of elevation. A detailed attribute comparison of the O-ESD dataset and the original Swiss Dwellings is brought in Table 3. Most of the attributes corresponding to the identifying and spatial categories were directly derived from the Swiss Dwellings dataset and have only undergone the refinement process, whereas the environmental attributes of O-ESD were the result of applying both refinement process and attribute transformation on the raw attributes in the Swiss Dwellings. The Venn diagram of the O-ESD attributes is presented in Figure 5, illustrating the overlap of some attributes in two or three categories. For instance, "zoning" can be simultaneously regarded as an identifying (e.g. differentiating bedrooms from dining spaces), a spatial (e.g. zone 1 being connected to zone 2) and an environmental attribute (e.g. high view quality and afternoon daylight being preferred in zone 2 areas). These sets of attributes enable performing tasks that require identification of spaces on various scales, combined with geometrical information and environmental performance proxy indicators.

The scatter pair-plot of the O-ESD's seven environmental attributes categorized by each zoning type is presented in Figure 6 to reveal insights regarding patterns in data and the pair-wise interplay among environmental factors. This plot is helpful to identify dependent and independent variables for further downstream tasks. On the diagonal of the plot, distribution of environmental data is shown by the kernel density estimate (KDE) method, across different zones. The plot shows a strong correlation between the morning and afternoon illuminance factors in every zone, which is expectable as both proxy indicators are measures of daylight illuminance at different times of the same day. Another positive correlation can be noticed between each of the illuminance factors and the view layer sky, showing more daylight access potential with more visible sky. Also, given that view layers sky and urban are negatively correlated, both of the illuminance factors show a negative correlation with the view layer urban. Compared to the other view layers (i.e. landscape, ground and sky) the view layer urban demonstrates a negative correlation. No apparent correlation between noise night and any

Table 2. The number of rows, columns and memory usage related to each step of data refinement and enrichment towards O-ESD dataset. Note that the percentage reductions are according to the previous step in the same data frame until the merging operation

Enrichment phase	Action steps (step number)	Data frame	Row count (percentage decrease)	Column count (percentage decrease)	Memory usage (percentage decrease)
Data refinement	Removing duplicates (01)	Simulation	367,678 (0%)	369 (0%)	1010.6 MB (0%)
	Removing duplicates (02)	Geometries	3,255,786 (0.0037%)	13 (0%)	322.9 MB (0%)
	Excluding non-residential unit usages (03)	Geometries	2,650,729 (18.5841%)	12 (7.6923%)	262.9 MB (18.5816%)
	Removing furniture-like features (04)	Geometries	2,385,165 (10.0185%)	12 (0%)	236.6 MB (10.0038%)
	Removing railings (05)	Geometries	2,239,274 (6.1166%)	12 (0%)	222.1 MB (6.1285%)
	Resetting index (06)	Geometries	2,239,274 (0%)	12 (0%)	205.0 MB (7.6992%)
	Maintaining related attributes (07)	Simulation	367,678 (0%)	36 (90.2439%)	101.0 MB (90.0059%)
Attribute transformation	Grouping to zoning (08)	Geometries	2,239,274 (0%)	13 (−8.3333%)	222.1 MB (−8.3414%)
	Adding illuminance factor proxies (09)	Simulation	367,678 (0%)	38 (−5.5556%)	106.6 MB (−5.5445%)
	Removing raw daylight attributes (10)	Simulation	367,678 (0%)	36 (5.2632%)	101.0 MB (5.2533%)
	Adding view layers (11)	Simulation	367,678 (0%)	40 (−11.1111%)	112.2 MB (−11.0891%)
	Removing raw view attributes (12)	Simulation	367,678 (0%)	23 (42.5%)	64.5 (42.5134%)
	Adding noise night attribute (13)	Simulation	367,678 (0%)	24 (−4.3478%)	67.3 MB (−4.3411%)
	Removing raw noise attributes (14)	Simulation	367,678 (0%)	22 (8.3333%)	61.7 (8.3209%)
	Merging data frames (15)	Geometries and Simulation*	2,949,762	29	652.6 MB
	Applying threshold on noise data (16)	Merged dataset	2,949,762 (0%)	30 (−3.4482%)	675.1 MB (−3.4477%)
Data refinement	Removing complicated layouts in terms of window numbers (17)	Merged dataset	2,920,286 (0.9992%)	30 (0%)	690.7 MB (−2.3107%)
	Removing complicated layouts in terms of door numbers (18)	Merged dataset	2,731,865 (6.4521%)	30 (0%)	646.1 MB (6.4572%)
Attribute addition	Adding window orientations (19)	Merged dataset	2,712,485 (0.7094%)	31 (−3.3333%)	641.5 MB (0.7119%)

Note(s): *The values of this row act as the reference for calculating the percentage decreases of the following step's operation on the merged dataset

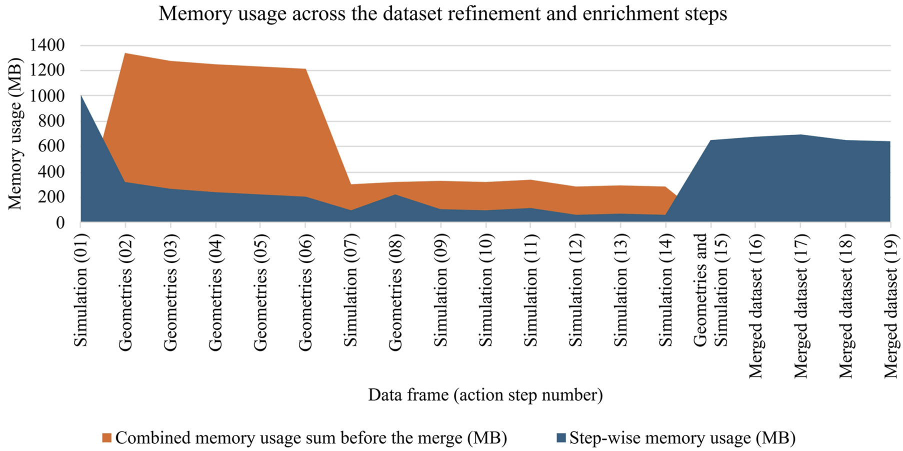

Figure 4. Memory usage across the dataset refinement and enrichment towards O-ESD

Table 3. O-ESD and Swiss dwellings attribute comparison

Attribute	Swiss dwellings	O-ESD
Site number	1,466	1,441
Building number	3,184	3,135
Apartment unit number	47,284	44,060
Room number	315,036	240,467
Floor range	−4 to 20	−4 to 20
Daylighting attributes number (description)	126 (minimum, p20, p80, mean, median, standard deviation, and maximum illuminance on the 21st of March, June, and December from 8:00 to 18:00 with a 2-h interval)	2 (morning and afternoon illuminance factor proxies on December 21st (overcast sky condition) at 10:00 and 16:00)
View attributes number (description)	126 (minimum, p20, p80, mean, median, standard deviation, and maximum view values of buildings, greenery, ground, isovist, mountain class 2, 3, 4, 5 and 6, railway tracks, site, sky, primary, secondary and tertiary streets, pedestrians, highways and water.)	4 (grouped view layers: urban, landscape, ground and sky)
Noise attributes number (description)	12 (minimum and maximum noise values of traffic, train, window noise traffic and window noise train for day and night)	2 (noise night (traffic and train), and acoustic comfort)
Window orientation	none	included (with 8 category precision)
Zoning	none	included (four zones according to the environmental demands)
Memory usage	1333.5 MB	641.5 MB

other environmental attributes can be observed. This can indirectly imply the effectiveness of the façade and window materials in blocking the outdoor noise, also resulting in more evenly distribution of noise levels across zones with prolonged occupancy (i.e. zone 1 and 2). All the illuminance factors and view layers expect for view layer urban demonstrate a right-skewed distribution in the KDE plots, showing that most data points are concentrated on the lower end,

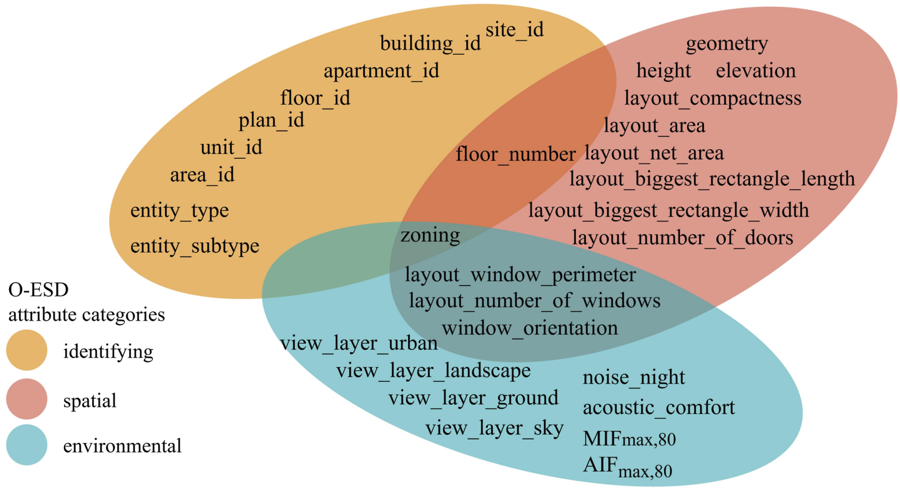

Figure 5. Venn diagram of the O-ESD attributes

but a few higher values stretch the distribution. The only left-skewed distribution belongs to view layer urban, with data being concentrated on higher values. All the indoor zones (i.e. zone 1,2 and 3) are mostly populated closely in the scatter pair-plots, and the extreme values mainly correspond to the outdoor zone 4. Among all of the zones, zone 3 appears to have significantly higher densities in lower values of natural environmental attributes (i.e. morning and afternoon illuminance factors, and view layers landscape, ground and sky), and higher values of one built environmental attribute (i.e. view layer urban). Moreover, zone 3 shows higher distribution density in the lower values of noise night data, which can be attributed to the fact that the entity subtypes of this zone (i.e. corridor, bathroom, storeroom and staircase) either do not have windows to the outdoor or are predominantly located far from the building boundaries.

As a result of adding the orientation attribute as the last step of the dataset enrichment, O-ESD contains 47,391 windows located towards North, 49,136 towards South, 48,721 towards East, 4,788 towards West, 5,148 towards North-East, 4,176 towards North-West, 3,565 towards South-East and 4,176 towards South-West. This knowledge of windows orientation in combination with the added environmental attributes and spatial zoning gives the opportunity for informative micro-climatic visualizations. An example is shown in Figure 7, where a floor layout is overlaid with window orientations, different environmental attributes and zonings in multiple steps. It can be observed that while the rightest room in zone 1 (bedroom zones) has the advantage of high view to greenery, it is the noisiest room during the night and does not show a high $MIF_{max,80}$ either. Therefore, placing the spaces belonging to zone 1 in that corner of the layout does not seem environmentally efficient. These visualizations with multiple overlaying possibilities offer a human-interpretable way of floor layout visual evaluation.

O-ESD is the first floor layout dataset with the window orientations explicitly as an attribute. Although the accuracy of window orientations is relatively high, there are some incorrectly labeled windows due to the procedure of orientation assignment. Multiple examples are brought in Figure 8. This matter mostly occurs in concave floor layouts (first and second examples in Figure 8), or when a vertical-shaped window is located on one side of the floor layout, but it is tilted towards the other side (third example in Figure 8). These exceptions must be carefully considered in further downstream tasks using O-ESD.

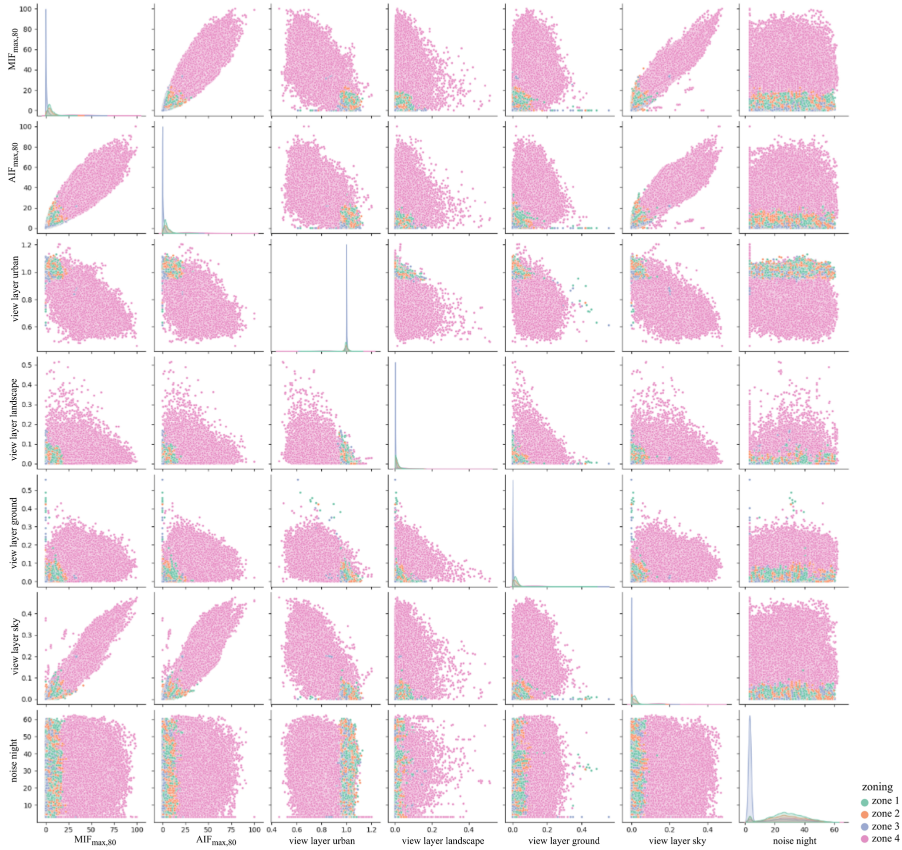

Figure 6. Scatter pair-plot of O-ESD environmental attributes with the diagonal KDE distribution

4.2 O-ESD limitations

Although O-ESD offers multiple application opportunities, there are some inherent limitations that need to be taken into account when employing the dataset. Depending on the downstream tasks, some limitations of the dataset may not be relevant; nevertheless, it is informative to be aware of the restrictions. The limitations can be categorized as follows:

(1) **Annotation style:** Inconsistencies among the annotation style of entity subtypes (i.e. room functions) among the Swiss Dwellings' floor layouts could be a cause of confusion in further implementation of O-ESD as well. For instance, some areas are annotated as "ROOM" (hence, not clearly a bedroom or living room) entity subtype, which could be practically a mixed functional combination of bedroom and living room. In O-ESD, these types of entity subtypes have been regarded as bedroom, that is zone 1 when grouping areas.

(2) (**Simulation condition of environmental data:** There is limited information about the simulation condition of environmental data in the Swiss Dwellings dataset and hence, the O-ESD is built upon the inevitable assumptions regarding daylighting, view and noise conditions. For instance, the height of surfaces on which the illuminance is simulated is unknown. Moreover, simulation conditions of the view and

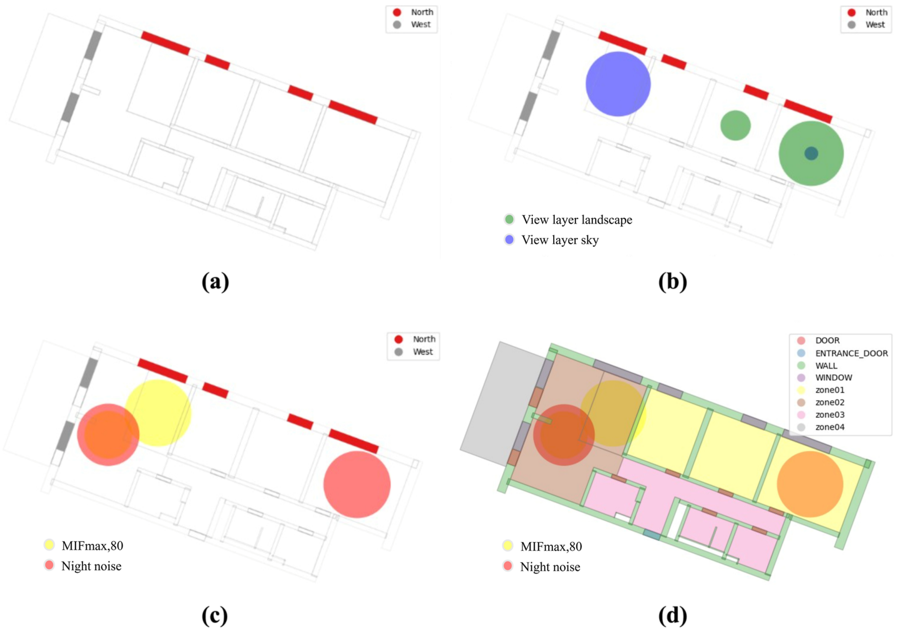

Figure 7. Different micro-climatic context visualizations of O-ESD data instances. (a) shows a floor layout marked with windows orientations, (b) overlays two selected view layers and marks the spaces with the highest values of them, (c) overlays the morning illuminance factor and a noise night proxy indicators and (d) represents the same environmental attributes shown in (c) but also overlaid on a zoning map

noise data are not comprehensively explained in the Swiss Dwellings dataset description.

(3) **Building material:** As a consequence of the Swiss Dwellings dataset's unavailability of data on building component materials, O-ESD also lacks this attribute. Since the type of material in the load-bearing structure, interior walls and ceilings and façade components play a central role in determining the operational energy consumption and embodied carbon footprint of the building, O-ESD would be highly limited in energy- and carbon-related study directions.

Tackling all the above-mentioned points would facilitate the implementation of O-ESD in various possible applications. Therefore, O-ESD should be cautiously applied in this regard.

4.3 O-ESD potential applications

As a bundle of identifying, spatial and environmental floor layout attributes, O-ESD dataset can facilitate downstream tasks in multiple ways. This helps integrating environmental aspects along with geometrical data on various scales from individual rooms to zones, apartment units, floors, buildings or sites. Among other possibilities, some applications for O-ESD are envisioned as follows:

(1) **Conditional floor layout generation:** In a conditional floor layout generation setting, the shape of a floor layout can be learned from the geometry data distribution, guided by defined environmental conditions such as orientation. Furthermore, design or evaluation constraints can be applied at different stages or different scales. For

Figure 8. Examples of floor layouts in O-ESD with an incorrectly labeled window orientation in the window orientation maps

instance, geometrical constraints such as building boundaries can appear earlier than zone-wise environmental demand evaluation.

(2) **Floor layout assessment:** Given the thresholds on environmental attributes in building certification systems and relevant standards, O-ESD floor layouts themselves or generated floor layouts using O-ESD layouts as training set can be assessed regarding aspects including visual (daylight provision and view quality) and acoustic comfort.

(3) **Floor layout similarity detection:** O-ESD provides several attributes with which the floor layouts can be compared against one another. A combination of geometrical (e.g. using plan id) and environmental (e.g. orientation of windows) attributes can help finding similar floor layouts based on an intended criterion. This would also be useful in the downstream tasks that deal with classification of floor layouts based on stereotypes using computer vision.

5. Conclusion

As a result of the shift towards data-driven methods, the significance of well-organized datasets containing comprehensive attributes is therefore more than ever on spotlight. In this regard, this study aimed to make a bond between architectural datasets and multiple typology and climate-aligned building performance guidelines by formalizing environmental architectural data into a structured, machine-readable and learnable framework. To the best of our knowledge, this work represents the first attempt to systematically structure environmental architectural clues into a unified dataset designed explicitly to support quantitative evaluation and AI-driven workflows. As the material for both ends of this bond, Swiss Dwellings dataset containing geometrical and environmental data of existing housing in Switzerland and BREEAM International New Construction, EN 17037, and Night noise guideline for Europe were selected. As a result of this bond, an environmentally enriched version of the Swiss Dwellings dataset was provided with a step-by-step data refinement and attribute transformation approach, leading to O-ESD dataset. The dataset enrichment process resulted in 10 new refined and transformed identifying, spatial and environmental attributes.

To leverage the existing data in architecture rather than building them from scratch, the clarification of the roadmap to O-ESD was also presented in this study to lay the ground for other possible dataset enrichments. Although the one-to-one mapping of the data refinement, attribute transformation and attribute addition approach might not be directly applicable for other potential datasets, the proposed roadmap offers promising and adjustable methodology. The examples of such adjustments include choosing various thresholds for identifying the complicated floor layouts and different daylight, view and noise proxy indicators according to underlying data availability and target tasks.

O-ESD potential ML-based downstream applications include conditional floor layout generation, floor layout assessment and similarity learning. For instance, the orientation attribute of O-ESD could serve as an input condition of a floor layout generation workflow, guiding the output layout towards a context-informed design. Given that the attributes offered by O-ESD are highly tailored to the schematic design phase requirements, the embedded environmental knowledge in the dataset could be jointly recognized with the floor layouts' geometrical features to enable the above-mentioned applications. While O-ESD contains precise additional attributes such as zoning types and orientations, the transformed environmental attributes including illuminance factors, view layers and noise night values act as guideline-inspired proxy indicator. When implementing O-ESD, its inherent limitations regarding annotation style, simulation condition for environmental data and lack of building material should be taken into account. The O-ESD dataset, the python code roadmap towards it and example visualizations are available through the github repository via this link https://github.com/Fatemeh-Mostafavi/O-ESD

Acknowledgments
Authors state that no human participants or animals were involved in this research. Moreover, no AI-assisted technologies were used for the writing purposes of the manuscript.

References

Abdelrahman, M., Coates, P. and Poppelreuter, T. (2023), "Visible outside view as a facilitation tool to evaluate view quality and shading systems through building openings", *Journal of Building Engineering*, Vol. 80, 108049, doi: 10.1016/J.JOBE.2023.108049.

Alsaggaf, F. (2024), "Dementia-inclusive design", TU Delft Repository, available at: https://repository.tudelft.nl/record/uuid:98b925fd-fd53-444f-a1f6-4863df303e3b (accessed 20 October 2025).

Amor, R. and Dimyadi, J. (2021), "The promise of automated compliance checking", *Developments in the Built Environment*, Vol. 5, 100039, doi: 10.1016/j.dibe.2020.100039.

Autodesk Forma (2025), "Forma site design", available at: https://www.autodesk.com/eu/products/forma/overview (accessed 31 October 2025).

Beach, T.H., Hippolyte, J.L. and Rezgui, Y. (2020), "Towards the adoption of automated regulatory compliance checking in the built environment", *Automation in Construction*, Vol. 118, 103285, doi: 10.1016/j.autcon.2020.103285.

Bielik, M., Zhang, L. and Schneider, S. (2023), "Big data, good data, and residential floor plans: feature selection for maximizing the information value and minimizing redundancy in residential floor plan data sets", *Communications in Computer and Information Science*, Vol. 1819 CCIS, pp. 607-622, Springer Science and Business Media Deutschland GmbH, doi: 10.1007/978-3-031-37189-9_40/TABLES/2.

BRE Global (2016), *BREEAM New Construction 2016*, BRE Global, Watford, available at: https://breeam.com/about (accessed 6 February 2026).

BRE Global (2021), *BREEAM International New Construction Version 6.0, BRE Global, Hertfordshire*, available at: https://breeam.com/standards (accessed 6 February 2026).

Building Research Establishment Ltd (2006), *Ecohomes 2006-The Environmental Rating for Homes*, BRE Global, Garston, April, available at: https://www.thenbs.com/PublicationIndex/documents/details?Pub=BRE&DocId=281258 (accessed 6 February 2026).

Chelaru, B., Onuțu, C., Ungureanu, G., Volf, I. and Şerbănoiu, A.A. (2025), "BIM and cloud-based tools integration for BREEAM in energy-efficient building design: a case study", *Results in Engineering*, Vol. 26, 105493, doi: 10.1016/J.RINENG.2025.105493.

Clark, C. and Paunovic, K. (2018), "WHO environmental noise guidelines for the European region: a systematic review on environmental noise and quality of life, wellbeing and mental health", *International Journal of Environmental Research and Public Health*, Vol. 15 No. 11, p. 2400, MDPI, doi: 10.3390/IJERPH15112400.

Croffi, J., Soebarto, V., Kroll, D. and Barrie, H. (2025), "A machine-learning approach for evaluating occupants' indoor environment satisfaction in high-rise mixed-use buildings", *Smart and Sustainable Built Environment*, Vol. ahead-of-print, pp. 1-30, doi: 10.1108/sasbe-03-2025-0161.

Dalux (2025), "Dalux | Software for the construction industry", available at: https://www.dalux.com (accessed 31 October 2025).

European Committee for Standardization (2021), "NEN-EN 17037: daylight in buildings + A1 annex", available at: https://www.nen.nl/en/nen-en-17037-2018-a1-2022-nl-291727 (accessed 6 February 2026).

Graaf, R.de. (2023), *Architect, Verb : The New Language of Building*, Verso, London, 9781839761935.

Greenberg, J., Wu, M.F., Liu, W. and Liu, F. (2025), "Metadata as data intelligence", *Data Intelligence*, Vol. 5 No. 1, pp. 1-5, doi: 10.1162/DINT_E_00212.

Hagedorn, P., Pauwels, P. and König, M. (2023), "Semantic rule checking of cross-domain building data in information containers for linked document delivery using the shapes constraint language", *Automation in Construction*, Vol. 156, 105106, doi: 10.1016/j.autcon.2023.105106.

Hosseini, S., Shabani, M.A., Irandoust, S. and Furukawa, Y. (2023), "PuzzleFusion: unleashing the power of diffusion models for spatial puzzle solving", *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Curran Associates, doi: 10.48550/arXiv.2211.13785.

Hosseini, S.M., Showkatbakhsh, M., Mahdavinejad, M. and Najafi, M. (2025), "An integrated generative-analytic framework for the performance-driven design of kinetic façades", *Smart and Sustainable Built Environment*, Vol. ahead-of-print, pp. 1-28, doi: 10.1108/SASBE-07-2025-0388/1301412.

Hraška, J. and Čurpek, J. (2024), "The practical implications of the EN 17037 minimum target daylight factor for building design and urban daylight in several European countries", Vol. 10 No. 1, p. e23297, doi: 10.1016/j.heliyon.2023.e23297.

Hu, R., Huang, Z., Tang, Y., Van Kaick, O., Zhang, H. and Huang, H. (2020), "Graph2Plan: learning floorplan generation from layout graphs", *ACM Transactions on Graphics*, Vol. 39 No. 4, Association for Computing Machinery, doi: 10.1145/3386569.3392391.

Kamsu-Foguem, B., Abanda, F.H., Doumbouya, M.B. and Tchouanguem, J.F. (2019), "Graph-based ontology reasoning for formal verification of BREEAM rules", *Cognitive Systems Research*, Vol. 55, pp. 14-33, doi: 10.1016/J.COGSYS.2018.12.011.

Kanna, K., Ait Lachguer, K. and Yaagoubi, R. (2022), "MyComfort: an integration of BIM-IoT-machine learning for optimizing indoor thermal comfort based on user experience", *Energy and Buildings*, Vol. 277, 112547, doi: 10.1016/J.ENBUILD.2022.112547.

Kat, C.J., Mostafavi, F., Brembilla, E. and Turrin, M. (2024), "Application of multimodal learning in daylight provision and view quality assessment of residential building layouts", *International Journal of Architectural Computing*, Vol. 22 No. 4, pp. 605-627, doi: 10.1177/14780771241286614.

Lannelongue, L., Grealey, J. and Inouye, M. (2020), "Green algorithms: quantifying the carbon footprint of computation", *Advanced Science*, Vol. 8 No. 12, doi: 10.1002/advs.202100707.

LIFULL Co. Ltd (2015), "Informatics research data repository [LIFULL HOME'S dataset]", available at: https://www.nii.ac.jp/dsc/idr/en/lifull/ (accessed 6 February 2026).

Luo, Z. and Huang, W. (2022), "FloorplanGAN: vector residential floorplan adversarial generation", *Automation in Construction*, Vol. 142, 104470, doi: 10.1016/J.AUTCON.2022.104470.

Lystbæk, M.S. (2025), "Machine learning-driven processes in architectural building design", *Automation in Construction*, Vol. 178, 106379, doi: 10.1016/J.AUTCON.2025.106379.

MacLeamy, P. (2004), "Collaboration, integrated information and the project lifecycle in building design, construction and operation", available at: https://www.scribd.com/document/474888363/Collaboration-Integrated-Information-and-the-Project-Lifecycle-pdf (accessed 18 July 2025).

Mostafavi, F. and Khademi, S. (2023), "Micro-Climate Building Context Visualization A pipeline for generating buildings' environmental context maps using numerical simulation data", *41st Conference on Education and Research in Computer Aided Architectural Design in Europe, ECAADe 2023*, pp. 9-18, available at: https://research.tudelft.nl/en/publications/micro-climate-building-context-visualization-a-pipeline-for-gener/ (accessed 6 February 2026).

Mostafavi, F., van Engelenburg, C., Khademi, S. and Vrachliotis, G. (2025), "Floor plan generation: the interplay among data, machine, and designer", *International Journal of Architectural Computing*, Vol. 23 No. 2, pp. 370-386, doi: 10.1177/14780771241290649.

Nauata, N., Chang, K.-H., Cheng, C.-Y., Mori, G. and Furukawa, Y. (2020), "House-GAN: relational generative adversarial networks for graph-constrained house layout generation", Proceedings, Part I, *Computer Vision – ECCV 2020: 16th European Conference*, Glasgow, UK, August 23-28, 2020, Springer-Verlag, pp. 162-177, doi: 10.1007/978-3-030-58452-8_10.

Nauata, N., Hosseini, S., Chang, K.-H., Chu, H., Cheng, C.-Y. and Furukawa, Y. (2021), "House-GAN ++ : generative adversarial layout refinement networks", doi: 10.48550/arXiv.2103.02574.

Office of the European Union (2024), *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/*

858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Publications Office of the European Union, Luxemburg, available at: http://data.europa.eu/eli/reg/2024/1689/oj (accessed 6 February 2026).

Pizarro, P.N., Hitschfeld, N., Sipiran, I. and Saavedra, J.M. (2022), "Automatic floor plan analysis and recognition", *Automation in Construction*, Vol. 140, April, 104348, doi: 10.1016/j.autcon.2022.104348.

Pourel, D., Ghaffarianhoseini, A., GhaffarianHoseini, A. and Naismith, N. (2025), "Integrated considerations of thermal comfort, daylight comfort and outdoor microclimate in medium density housing (MDH) development", *Smart and Sustainable Built Environment*, Vol. ahead-of-print, pp. 1-25, doi: 10.1108/SASBE-05-2025-0286/1301974.

Raw, G.J. and Prior, J.J. (1991), *BREEAM/New Homes: The BRE Environmental Assessment Method for New HOMES, BRE Global, Garston*, available at: https://breeam.com (accessed 6 February 2026).

Reich, Y. (1997), "Machine learning techniques for civil engineering problems", *Computer-Aided Civil and Infrastructure Engineering*, Vol. 12 No. 4, pp. 295-310, doi: 10.1111/0885-9507.00065.

Sambasivan, N., Kapania, S. and Highfll, H. (2021), "Everyone wants to do the model work, not the data work: data cascades in high-stakes ai", *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, doi: 10.1145/3411764.3445518.

Satellite climatology - MeteoSwiss (2025), available at: https://www.meteoswiss.admin.ch/climate/the-climate-of-switzerland/satellite-climatology.html (accessed 4 July 2025).

Shabani, M.A., Hosseini, S. and Furukawa, Y. (2023), "HouseDiffusion: vector floorplan generation via a diffusion model with discrete and continuous denoising", *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, pp. 5466-5475, doi: 10.1109/CVPR52729.2023.00529.

Shen, Y. and Pan, Y. (2023), "BIM-supported automatic energy performance analysis for green building design using explainable machine learning and multi-objective optimization", *Applied Energy*, Vol. 333, 120575, doi: 10.1016/J.APENERGY.2022.120575.

Shim, J., Moon, J., Kim, H. and Hwang, E. (2024), "FloorDiffusion: diffusion model-based conditional floorplan image generation method using parameter-efficient fine-tuning and image inpainting", *Journal of Building Engineering*, Vol. 95, 110320, doi: 10.1016/J.JOBE.2024.110320.

Sobhkhiz, S. and El-Diraby, T. (2023), "Dynamic integration of unstructured data with BIM using a no-model approach based on machine learning and concept networks", *Automation in Construction*, Vol. 150, 104859, doi: 10.1016/J.AUTCON.2023.104859.

Standfest, M., Franzen, M., Schröder, Y., Gonzalez Medina, L., Villanueva Hernandez, Y., Hendrik Buck, J. and Tan, Y.-L. (2023), "Swiss dwellings", doi: 10.5281/zenodo.7788422.

Stouffs, R., Kooistra, J. and Tuncer, B. (2004), "Metadata as a means for correspondence on digital media", *Journal of Information Technology in Construction*, available at: https://www.itcon.org/paper/2004/9 (accessed 6 February 2026).

Sucholutsky, I., Battleday, R.M., Collins, K.M., Marjieh, R., Peterson, J.C., Singh, P. and Bhatt, U. (2023), "On the informativeness of supervision signals", *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pp. 2036-2046, doi: 10.48550/arXiv.2211.01407.

Sun, Y., Wu, Y. and Wilson, R. (2018), "A review of thermal and optical characterisation of complex window systems and their building performance prediction", *Applied Energy*, Vol. 222, pp. 729-747, doi: 10.1016/J.APENERGY.2018.03.144.

Sun, J., Wu, W., Liu, L., Min, W., Zhang, G. and Zheng, L. (2022), "WallPlan: synthesizing floorplans by learning to generate wall graphs", *ACM Transactions on Graphics*, Vol. 41 No. 4, pp. 1-14, Association for Computing Machinery, doi: 10.1145/3528223.3530135.

Tang, H., Zhang, Z., Shi, H., Li, B., Shao, L., Sebe, N. and Timofte, R. (2023), "Graph transformer GANs for graph-constrained house generation", doi: 10.48550/arXiv.2303.08225.

Upadhyay, A., Dubey, A., Arora, V., Kuriakose, S.M. and Agarawal, S. (2022), "FLNet: graph constrained floor layout generation", *ICMEW 2022 - IEEE International Conference on Multimedia and Expo Workshops 2022, Proceedings*, Institute of Electrical and Electronics Engineers, doi: 10.1109/ICMEW56448.2022.9859350.

van den Burg, G.J.J., Nazábal, A. and Sutton, C. (2019), "Wrangling messy CSV files by detecting row and type patterns", *Data Mining and Knowledge Discovery*, Vol. 33 No. 6, pp. 1799-1820, doi: 10.1007/s10618-019-00646-y.

van Engelenburg, C., Mostafavi, F., Kuhn, E., Jeon, Y., Franzen, M., Standfest, M. and van Gemert, J. (2025), *MSD: A Benchmark Dataset for Floor Plan Generation of Building Complexes*, Springer, Cham, pp. 60-75, doi: 10.1007/978-3-031-73636-0_4.

Vrachliotis, G. (2022), *Architecture and Design in the Age of Cybernetics*, Birkhäuser, Berlin, 9783035624816.

Waters, B.L., Dixit, M.K. and Pariafsai, F. (2023), "Architects' ranking of professional design services", *Architectural Engineering and Design Management*, Vol. 19 No. 6, pp. 620-641, doi: 10.1080/17452007.2022.2147896.

Weber, R.E., Mueller, C. and Reinhart, C. (2022), "Automated floorplan generation in architectural design: a review of methods and applications", *Automation in Construction*, Vol. 140, 104385, doi: 10.1016/j.autcon.2022.104385.

WHO (2009), *Night Noise Guidelines for Europe*, Regional Office for Europe, Copenhagen, available at: https://iris.who.int/handle/10665/326486 (accessed 6 February 2026).

Wu, W., Fu, X.M., Tang, R., Wang, Y., Qi, Y.H. and Liu, L. (2019), "Data-driven interior plan generation for residential buildings", *ACM Transactions on Graphics*, Vol. 38 No. 6, pp. 57-74, doi: 10.1145/3355089.3356556.

Zawidzki, M. and Szklarski, J. (2020), "Multi-objective optimization of the floor plan of a single story family house considering position and orientation", *Advances in Engineering Software*, Vol. 141, February 2019, 102766, doi: 10.1016/j.advengsoft.2019.102766.

Zhang, H. and Zhang, R. (2025), "Generating accessible multi-occupancy floor plans with fine-grained control using a diffusion model", *Automation in Construction*, Vol. 177, 106332, doi: 10.1016/j.autcon.2025.106332.

Zhong, B., Gan, C., Luo, H. and Xing, X. (2018), "Ontology-based framework for building environmental monitoring and compliance checking under BIM environment", *Building and Environment*, Vol. 141, pp. 127-142, doi: 10.1016/j.buildenv.2018.05.046.

Zhou, P. and El-Gohary, N. (2021), "Semantic information alignment of BIMs to computer-interpretable regulations using ontologies and deep learning", *Advanced Engineering Informatics*, Vol. 48, 101239, doi: 10.1016/j.aei.2020.101239.

Zhuang, X., Zhu, P., Yang, A. and Caldas, L. (2025), "Machine learning for generative architectural design: advancements, opportunities, and challenges", *Automation in Construction*, Vol. 174, 106129, doi: 10.1016/J.AUTCON.2025.106129.

Corresponding author
Fatemeh Mostafavi can be contacted at: f.mostafavi@tudelft.nl