Journal of Sustainability (2026) 2:1, 1-38
https://doi.org/10.55845/jos-2026-21101

RESEARCH ARTICLE

Determinants of Building-Sector CO_2 Emissions in the EU: A Combined Econometric and Machine Learning Approach

Marco Mele[1], Alberto Costantiello[2], Fabio Anobile[2], Angelo Leogrande[2*]

Received: 13. December 2025 / Accepted: 12. February 2026 / Published: 16. February 2026

Abstract

The research aims to explore the structural, environmental, and climatic factors that influence carbon dioxide emissions in the building sector across the 27 member states of the European Union between 2005 and 2023. It is evident that higher activity in the primary sector, along with higher forest area, is associated with reduced building sector emissions, while environmental stress, air pollution, and demands for heating and cooling are positively related to building sector emissions. Subsequent analyses show that there are significant differences between EU member states, differentiating between those that are classified under high-emitting patterns, including pollution, inefficient energy, and adverse climatic conditions, and those that are classified under low-emitting patterns, including clean energy sources and favourable environmental conditions.

Keywords Building-sector Carbon Emissions · Panel Data Econometrics · Machine Learning Prediction · Environmental and Climatic Drivers · Cluster Analysis

JEL Codes C33 · Q54 · Q41 · Q56 · C38

1. Introduction

The building industry is a major area of policy engagement in the European Union's decarbonisation plan, accounting for a substantial share of total energy use and greenhouse gas emissions (Giannelos et al., 2023). Although interest in building energy performance has been growing, knowledge of the structural, environmental, and climatic factors influencing the magnitude of CO_2 emissions in this area remains fragmented. Existing research usually focuses on separate variables—energy efficiency, electricity supply composition, and renovation strategies—without adopting a multidimensional approach to identify cross-correlations among economic structures, environmental conditions, climatic factors, and technology typologies in this area of analysis (Bezić et al., 2022). Furthermore, comparative analyses for a total of all member states of a political union like the European Union usually involve traditional econometric modelling that has not been coupled with contemporary machine learning models, potentially able to identify nonlinear patterns in available data analysis in this area of analysis as well (Giannelos & Bellizio, 2024; Sadhukhan & Yadav, 2023; Doran et al., 2025). However, this study fills this research gap by providing a comprehensive, systemic analysis of the factors influencing CO_2 emissions in the building sector across 27 EU Member States from 2005 to 2023. With this research gap in mind, this research is based on the following research question: What structural, environmental, and climatic drivers of CO_2 emissions exist for the building sector among European Union Member States, and how do these drivers interact when viewed through the lens of an econometric, machine-learning, and clustering model?

[1] Unicusano University, Rome
[2] LUM University Giuseppe Degennaro, Casamassima
[*] Corresponding Author: leogrande.cultore@lum.it

In order to break down this broad research question, this research will attempt to answer the following specific research questions: (I) How do environmental and climatic variables such as water stress, forest area, heating and cooling degree days, and air pollution influence CO_2 emissions for the building sector among European Union Member States? (II) What is the influence of agricultural and structural economic variables on the differences among European Union Member States for building sector CO_2 emissions? (III) Can nonlinear relationships and groupings uncovered through machine-learning and clustering analysis validate or support the results uncovered through traditional panel econometric models? Through this research, all of the above research questions will be addressed and answered.

The novelty of this study consists in applying a set of four panel data models—Random Effect Model, Fixed Effect Model, Dynamic Panel GMM Model, and Weighted Least Squares Model—simultaneously with a robust set of models using machine learning algorithms and a multi-layered cluster analysis, as in Cialani Mortazavi (2021. The simultaneous application of econometric models, which help understand causal effects and temporal variation, and machine learning models, which help understand nonlinear effects, overcomes issues related to using a single analysis model, as described in Vagnini (2025). On the other hand, Clustering analysis has been employed to understand country-level structural diversity, as described in Tudor (2025. This approach of using a set of models for analysis has been found to be very innovative, as very few studies, as cited in Giannelos (2023) and Giannelos & Bellizio (2024), used a set of models for analysis of emissions in the building sector of Europe.

This research also adds value by identifying a wide range of variables covering climatic, environmental, economic, and agricultural aspects comprehensively (Bezic et al., 2022; Doran et al., 2025). Variables such as heating and cooling degree days, particulate matter concentration ($PM_{2.5}$), total water stress, forest area, and nitrous oxide emissions are incorporated alongside economic and agricultural variables, such as agriculture, forestry, and fishing value added (AFFV) and the food production index. These are rarely quantified together in studies analysing building-related emissions, but all are important for understanding how combinations of land use, natural carbon absorption, environmental pressures, and economic systems relate to energy use in buildings and their emissions (Charlier et al., 2023).

Incorporation of agricultural variables in this study also fills a conceptual gap, as this set of variables has traditionally been linked only to agricultural emissions, but this study shows that it also applies to emissions from buildings (Vagnini et al., 2025; Cialani & Mortazavi, 2021). A further original aspect concerns the robustness analysis of models using empirical evidence of their predictive capabilities, as described in Giannelos et al. (2023). According to the econometric analysis, Fixed Effects and Weighted Least Squares models provide superior diagnostic performance, whereas Dynamic Panel GMM performance is affected by a lack of instrument validity, which has seldom been analysed in depth in a similar study (Bezić et al., 2022).

Meanwhile, from a machine learning analysis, it emerges that K-Nearest Neighbors stands out as a superior model for a predictive task, even when compared with models of a linear nature and Tree-based models, as in models of Random Forest and Decision Trees, as described in Giannelos and Bellizio in 2024, as well as in other studies in 2023, as in a paper from 2025, in a paper by Giannelos et al. (2023), a result which means that similarity patterns in a structure of building-related emissions in the EU are exceptionally localized. Another innovative aspect of this study is its ten clusters, identified using statistical methods such as the Bayesian Information Criterion, which are also quantified using silhouette values (Cialani and Mortazavi, 2021).

The clusters obtained help identify vastly different structural and environmental characteristics among EU nations. The clusters of high emissions are characterised by factors such as agricultural intensity, pollution, and low energy efficiency, which point to structural and environmental pressures driving high emissions. Low-emitting clusters, on the other hand, are characterised by high inclusion of renewable energy, low pollutant concentrations, favourable climatic conditions, and high forest cover (Koengkan et al., 2022). This study adds to the literature by providing a typology of emission patterns that reflects the diversity of European circumstances and could form a basis for different policies in this area (Vagnini et al., 2025; Doran et al., 2025).

In sum, this research contributes in several ways to the scientific debate about energy and emission issues. This research offers a completely novel degree of methodological integration between econometric analysis and prediction models; it enlarges the set of analysed variables in a comprehensive fashion, including climatic, environmental, and agricultural ones; it provides a stringent statistical validation of model performance; and it discerns characteristic structural patterns for EU member states that are immediately policy-relevant for energy and climate policies (Giannelos et al., 2023; Bezić et al., 2022; Cialani & Mortazavi, 2021).

The results obtained show that to provide a comprehensive understanding of the determinants of building-sector-related CO_2 emissions, it is not sufficient to rely solely on analytical models. Instead, a unitary,

comprehensive perspective must be adopted that considers all intersections and interactions among factors related to economic structures, environmental features, climatic constraints, and energy-system typologies (Tudor et al., 2025). This seems to be exemplified in this paper, which includes all the mentioned features, with a note on methodological integration as a viable, fruitful pathway for further research in this area.

This research aims to fill a number of gaps in current knowledge on CO_2 emissions in the building sector in the European Union by contributing to the current body of knowledge in a number of ways. First, while current knowledge is largely focused on specific factors or a set of countries, there is a lack of studies at an EU-wide scale, considering a long-term period in a recent context. Second, while current knowledge is largely based on a single approach to analysis, there is a lack of application of panel econometric models in combination with machine learning models and clustering analysis in current knowledge on CO_2 emissions in the building sector in the European Union. In light of these considerations, the purpose of this research is to provide a common framework to analyse how different characteristics, pressures, and conditions are associated with CO_2 emissions in the building sector in each of the 27 Member States of the European Union between 2005 and 2023.

1.1. Research Gap and Contribution of the Study

While the introduction provides the policy background, motivation, and objectives of the study, the literature review systematically examines recent empirical and methodological contributions on building-sector emissions. The present study addresses gaps identified in the literature by offering an EU-wide, multi-method analysis over a long and recent time horizon, integrating panel econometrics, machine learning, and clustering techniques. The specific objectives are to identify key structural, environmental, and climatic factors associated with building-sector CO_2 emissions and to explore heterogeneity across EU member states.

The article continues as follows: the second section reviews the literature on determinants of building-sector emissions. The third section presents the multi-method framework combining econometrics, machine learning, and clustering. The fourth section details the panel data models used to analyse emission drivers. The fifth section evaluates machine-learning performance and identifies KNN as the best predictor. The sixth section compares clustering algorithms and justifies the hierarchical model selection. The seventh section integrates findings from all methods to provide unified insights. The eighth section outlines the key policy implications derived from the results. The ninth section concludes the study.

2. Literature Review

In this literature review, a structured and open search process using the Scopus database and Google Scholar search engine will be adopted. The literature will be restricted to published articles in peer-reviewed journals between 2021 and 2025, using carefully identified keywords related to emissions in the building sector, energy efficiency, effects of climate change, econometric models, and machine learning techniques. The articles will be presented through clear thematic strands to avoid any repetition of the introduction section, which will be restricted to background information, motivations for conducting the research, and research aims.

More recent literature on decarbonisation in the building sector focuses on the fact that emissions are generated by a combination of structural, environmental, climatic, technological, and policy-related factors. Rather than solely attributing the dynamic of emissions to technological substitution, an increasing number of studies apply multi-dimensional conceptual frameworks in order to account for the complexity of emissions in buildings.

A first line of literature focuses on structural, economic, and geographic determinants of building-related emissions. Magaletti et al. (2025) argue that in order to reduce emissions, it is necessary that there be a structural shift in economic structure in tandem with technological advancement. By applying bottom-up modelling, Sarıca et al. (2023) show that structural composition and structural definition are crucial in assessing routes of reduced emissions in European economies. Within the EU, Tzeiranaki et al. (2023) show that differences in tertiary activity, climate, and structural differences have a significant effect on energy consumption patterns in buildings. Econometrically, complementary evidence comes from nonlinear STIRPAT models by Zhu et al. (2022), who show that economic growth, resource pressure, and industrialisation are major determinants of embodied emissions in construction, while emphasising spatial spillovers by cross-country panel models by Wei et al. (2023).

A second strand of literature focuses on materials, typologies, and embodied carbon in buildings. Rheude & Röder (2022) show that materials are a major source of greenhouse gas emissions in the German building stock. Related literature underlines that there are decarbonisation potentials in alternative materials in construction, such as wood and timber (Reyes et al., 2021; Piccardo et al., 2021), while Hemmati et al. (2024), by contrast, show that design matters in embodied emissions in steel-and concrete-based buildings. Data-driven assessments of renovation and demolition options further show that renovation-oriented options are superior in environmental terms compared to demolitions (Famiglietti et al., 2023; Dragonetti et al., 2025).

A third line of literature focuses on policy tools and frameworks. Attia et al. (2021), by econometric analysis, find that sustainability certification, such as LEED, can contribute significantly to reduced emissions if properly applied. Braungärdt et al. (2025), by assessing the possible extension of an emissions trading system in the EU construction industry, show that while carbon pricing can contribute to decarbonisation, additional measures are necessary in order to mitigate distributional effects. Kartal et al. (2024) further illustrate that the relationship between environmental policy stringency and emissions differs strongly across EU countries.

More recently, an increasing number of studies have applied data-intensive and machine learning techniques to analyse building sector emissions. Giannelos et al. (2024) illustrate that nonlinear machine learning models outperform standard models in forecasting building-related CO_2 emissions, whereas Cubuk (2023) finds that there are large differences between ex-ante and ex-post emissions. Machine learning techniques are also applied to analyse stakeholder behaviour and future infrastructure/material trends. Rakhshan et al. (2023) analyse stakeholder behaviour in this way, whereas Gan et al. (2023) and Papachatzis (2024) analyse future infrastructure and material trends. The studies support the combination of econometric and data-intensive techniques.

Other studies focus on the topics of finance and circular economy strategies as well as on sociocultural aspects. Hszholipour et al. (2022) illustrate that green finance helps to reduce emissions in the construction sector, whereas Sharmina et al. (2023) argue that circular economy strategies are crucial for achieving net-zero goals. Qualitative studies emphasise that resistance and governance fragmentation are crucial barriers to decarbonisation in the building sector. Heinz et al. (2025) and Chen et al. (2024) illustrate this point.

Finally, studies on methodological aspects focus on integrated and transdisciplinary approaches. The studies by Mandel et al. (2023), Xin et al. (2023), Myint et al. (2025), and Hu et al. (2025) investigate ways toward carbon-neutral buildings. The studies by Li et al. (2024) and Papangelopoulou et al. (2025) investigate decomposition analysis and retrofit assessment.

The studies on energy flexibility and renovation effectiveness illustrate the interplay between technological, structural, and institutional aspects. Vigna et al. (2021) and Kadrić et al. (2022) illustrate this point in this way, whereas Żelazna and Pawłowski (2025) focus on energy flexibility and effectiveness in this way. The studies indicate that there is a need for an integrated and multi-country empirical analysis framework for capturing environmental pressure, economic structure, and technological processes. However, existing studies often appear to be fragmented in terms of their analysis techniques and geographical focus. The studies often focus on long-term EU-wide analysis and therefore motivate this study's integrated empirical analysis framework that combines panel econometrics and machine learning techniques to analyse the determinants of CO_2 emissions in the European building sector. See Table 1.

Table 1. Overview of Macro-Themes, Methodologies, and Main Results in the Building-Sector Decarbonisation Literature.

Thematic Area	Key References	Predominant Methods	Core Insights Relevant to This Study
Structural, Environmental, and Economic Drivers of Building-Sector Emissions	Magaletti et al. (2025); Sarıca et al. (2023); Tzeiranaki et al. (2023); Zhu et al. (2022); Erdogan (2021); Wei et al. (2023); Yakymchuk & Rataj (2025); Gan et al. (2023); Xin et al. (2023); Xia et al. (2024)	Panel and spatial econometrics, STIRPAT models, scenario analysis, bottom-up and macroeconomic modelling	Emissions are driven by structural economic characteristics, climatic conditions, technological change, and spatial spillovers. Innovation and structural transitions reduce long-run emissions, while scenario models identify heterogeneous pathways toward emission peaks and neutrality.

Note: This table organises recent literature into coherent thematic areas to clarify the theoretical and empirical foundations underlying the selection of variables and methods adopted in this study. By linking dominant research themes with methodological approaches and key findings, the table reduces fragmentation in the literature and provides a transparent rationale for the ESG-based and multi-method empirical framework employed.

Table 1 (Cont.). Overview of Macro-Themes, Methodologies, and Main Results in the Building-Sector Decarbonisation Literature.

Thematic Area	Key References	Predominant Methods	Core Insights Relevant to This Study
Policies, Governance, and Regulatory Instruments	Attia et al. (2021); Mandel et al. (2023); Braungärdt et al. (2025); Kartal et al. (2024); Cubuk (2023); Boca et al. (2025); Heinz et al. (2025)	Policy evaluation, ex-ante regulatory modelling (ETS), qualitative case studies, quantile regressions	Governance quality and policy design strongly condition decarbonisation outcomes. Carbon pricing and efficiency policies reduce emissions but require complementary measures to address distributional effects, implementation gaps, and sociopolitical barriers.
Structural, Environmental, and Economic Drivers of Building-Sector Emissions	Magaletti et al. (2025); Sarıca et al. (2023); Tzeiranaki et al. (2023); Zhu et al. (2022); Erdogan (2021); Wei et al. (2023); Yakymchuk & Rataj (2025); Gan et al. (2023); Xin et al. (2023); Xia et al. (2024)	Panel and spatial econometrics, STIRPAT models, scenario analysis, bottom-up and macroeconomic modelling	Emissions are driven by structural economic characteristics, climatic conditions, technological change, and spatial spillovers. Innovation and structural transitions reduce long-run emissions, while scenario models identify heterogeneous pathways toward emission peaks and neutrality.
Policies, Governance, and Regulatory Instruments	Attia et al. (2021); Mandel et al. (2023); Braungärdt et al. (2025); Kartal et al. (2024); Cubuk (2023); Boca et al. (2025); Heinz et al. (2025)	Policy evaluation, ex-ante regulatory modelling (ETS), qualitative case studies, quantile regressions	Governance quality and policy design strongly condition decarbonisation outcomes. Carbon pricing and efficiency policies reduce emissions but require complementary measures to address distributional effects, implementation gaps, and sociopolitical barriers.
Materials, Embodied Carbon, and Circular Economy Strategies	Reyes et al. (2021); Rheude & Röder (2022); Aste et al. (2022); Asdrubali et al. (2025); Chen et al. (2024); Seyedabadi et al. (2024); Hemmati et al. (2024); Hassan & Rezaei (2025); Dragonetti et al. (2025); Almusaed et al. (2024); Junda & Málaga-Chuquitaype (2025)	Life-cycle assessment (LCA), material flow analysis, structural and design comparisons, circular economy evaluation	Material choice and construction design significantly affect embodied emissions. Timber, recycled materials, and deep renovation strategies outperform demolition-based approaches, while regulatory standards and deterioration dynamics shape long-term carbon performance.
Modelling Tools, Machine Learning, and Technological Innovation	Giannelos et al. (2024); Rakhshan et al. (2023); Papachatzis (2024); Famiglietti et al. (2023); Meena et al. (2022); Papangelopoulou et al. (2025); Vigna et al. (2021); Kadrić et al. (2022); You et al. (2025)	Machine learning, spatially resolved LCA tools, energy-flexibility models, cost-effectiveness and innovation analysis.	Data-driven methods capture nonlinearities and improve predictive accuracy. Energy flexibility, demand-side management, and innovation-oriented tools support net-zero pathways and inform more effective building-sector policies.

Note: This table organises recent literature into coherent thematic areas to clarify the theoretical and empirical foundations underlying the selection of variables and methods adopted in this study. By linking dominant research themes with methodological approaches and key findings, the table reduces fragmentation in the literature and provides a transparent rationale for the ESG-based and multi-method empirical framework employed.

3. A Multi-Method Framework for Modelling the Building Sector CO_2 Emissions in the European Union

Due to the intricacies involved in the drivers of CO_2 emissions in the buildings sector, this study uses an extensive multi-method empirical approach in order to incorporate structural, environmental, climatic, and energy-related inter-connections. The concurrent multidimensional strategies have already proven successful in contemporary European studies on regional decarbonisation and socio-environmental drivers of emissions (Vagnini et al., 2025).

The goal in this study will be to investigate how environmental determinants, economic and structural attributes, and energy system attributes cumulatively shape individual and aggregate levels of CO_2 emissions

in buildings for Member States in the European Union. The mean effect of such variables will be estimated through panel data econometric models that account for country-level heterogeneity and shared time trends. The application of such econometric models as Fixed Effects and Dynamic Panel GMM estimators stems from recent empirical studies that have ascertained their appropriateness for environmental and energy-related emission analysis (Papadas et al., 2024).

The variables that represent environmental and climate factors are those that measure forest area, water stress, air pollution, and both heating and cooling degree days. Economic and structural factors are represented by variables that measure land use and production structure through agricultural value added and the food production index. The characteristics of energy systems and emission indicators are those that are indirect measures of variation in policy performance, technological diffusion, and compliance with environmental regulations within the context of the EU institutional setting.

A total of four panel equation methods are explored and employed in the empirical analysis. Tests of diagnostics reveal that Fixed Effects and Weighted Least Squares methods are superior, although Dynamic Panel GMM methods are also included to address potential issues of persistence and endogeneity, although with some reservations concerning their instruments (Özen et al., 2025). This approach to combining both static and dynamic panel methods is consistent with current empirical trends that seek to combine several methods of analysis to improve robustness and performance when studying emissions (Özen et al., 2025).

Regression machine-learning algorithms are used as a complement to better identify nonlinear relationships and complex dynamics not easily accounted for by traditional parametric econometric models. Machine-learning algorithms improve forecast accuracy and allow a better understanding of emission dynamics, reflecting a new trend to combine machine learning and traditional econometric analysis for research on environmental and energy economics (Drago et al., 2025).

Lastly, clustering analysis is used to identify Member States sharing structural, environmental, and climatic features, allowing a typological interpretation of emission dynamics. This method helps to identify a set of homogenous countries and to interpret differentiated dynamics of building sector emissions in the European Union.

3.1. Variable selection

The empirical approach uses three different specifications of models to identify the determinants of CO_2 emissions in the buildings sector. While one focuses on environmental factors, another focuses on structural factors. The third focuses on energy system-related factors. The use of a modular approach in modelling helps to identify the contribution of each determinant of CO_2 emissions in the buildings sector. The use of a similar modular approach has been identified in recent empirical research that has explored the relationship between emissions, effectiveness of policies, and sustainability (Iwanicz-Drozdowska et al., 2025).

Each variable in the specifications has been chosen on the basis of the most up-to-date literature in the field of economics and energy studies. Environmental and climate variables include the extent of forests, water stress, particulate matter concentration (PM2.5), and the number of heating and cooling degree days, recognised in the literature as important drivers of energy demand and environmental pressure. The structural and economic variables describe land use, the structure of the economy, and production structures, such as the value added in the agriculture, forestry, and fishing industries, as well as the food production index, chosen as they indirectly describe structural properties, gaining increased focus in recent cross-country studies (Nurgaliyeva, 2025; Costantiello et al., 2025).

The third specification focuses on the variables associated with the energy system and emissions, taking into account the variations in the level of energy efficiency, fuel types, and the level of compliance with environmental norms in the EU framework. Although the governance variables specific to the sector in the case of buildings cannot be considered because of the limitations in the data at the EU level, the chosen variables associated with the energy system and emissions are widely used in the literature as proxies for the performance of policies in the sector (Mohammed et al., 2024; Özen et al., 2025).

All the variables that are incorporated into the analysis are backed up with an ex-ante rationale, which is both theoretical and empirical, and is in line with the best practices for the application of econometric analysis on the subject of emissions and climate change policies (Costantiello et al., 2025; Iwanicz-Drozdowska et al., 2025). There is a refinement procedure that is utilised for the purpose of robustness analysis, which is used for the evaluation of the sensitivity of the outcomes and for the selection of the variables that have less explanatory value (Aquino et al., 2024; Özen et al., 2025).

A multi-method approach to analysis that combines panel econometric modelling, machine learning regressions, and cluster analysis across 27 Member States of the European Union has been adopted for research into the factors influencing CO_2 emissions in the construction sector (Giannelos et al., 2023; Bărbulescu, 2024). The multi-method technique adopted addresses various factors and requires tools capable of providing appropriate insights into specific areas associated with either direct causality or nonlinear prediction paths (Yang et al., 2023). A single methodological approach cannot deal with these complexities. As a result, the discussion will be structured around three methodological pillars.

The first pillar includes panel econometric models. The models use Fixed Effects, Random Effects, Dynamic Panel GMM, and Weighted Least Squares methods to identify statistically significant relationships, while adjusting for unobserved factors, persistence, and heteroscedasticity in the emission variables (Bărbulescu, 2024). Despite their traditional use in panels, instrumental variables have been deliberately avoided in this research. Within diffuse, geographically interdependent realms of environmental and climate affairs, with cross-regionally and transboundary processes and interactions, it would be extremely difficult to identify valid, exogenous variables suitable for use in instrumental variable estimation procedures (Yang et al., 2023). Variables such as temperature variability, emission intensities, and resource scarcity are endogenous and jointly determined within regions; thus, strict exogeneity would be empirically impractical.

The second pillar relates to machine learning via regression. Machine learning regressions assist and supplement econometric modelling by capturing nonlinear interrelations and interactions beyond parametric specifications. Machine learning regressions are particularly useful for detecting drivers and discovering hidden patterns in large environmental datasets. Equilibrium between econometric predictions and machine learning forecasts enhances confidence levels, but discrepancies can yield insights into dynamic forces suggested by machine learning outputs and warrant theoretical and econometric examination (Giannelos et al., 2023).

The third methodological pillar is clustering analysis. Due to considerable variability across EU Member States in climate, energy structures, and socioeconomic variables, clustering methods are applied to group Member States based on similarities in emission and structural patterns (Morelli et al., 2025; Sechi et al., 2022). A solution on a ten-cluster basis, as determined by the Bayesian Information Criterion and Silhouette analysis, helps achieve efficient stratification between low- and high-emission regimes. It should be remembered that these stratified policy prescriptions are necessary and efficient measures as compared to common prescriptions that disregard structural divergences among various nations (Morelli et al., 2025).

The methodological framework is therefore supported by the focus on the EU-27 in the analysis. Although small-N panel designs pose challenges regarding overfitting and testing power, there are nonetheless valid reasons for diversification within these designs (Bărbulescu, 2024). Both traditional econometric and machine learning techniques remain valid within these designs, provided they are properly tested and verified (Giannelos et al., 2023). Specifically, clustering analysis becomes more valid with medium-sized samples, as there would be fewer dangers associated with over-fragmentation as well as label ambiguity (Morelli et al., 2025). Third, the European Union is a particularly fruitful empirical case. It features a relatively homogeneous set of rules and, at the same time, considerable variation in weather, economic systems, and structure. It thus serves as an optimal testing ground and requires careful interpretation when assessing implications for global realities.

For all econometric models done in this study, data management, estimation, and model diagnostics were done using Stata software. Both machine-learning regression models and clustering models were done in JASP software, which provides a clear framework for supervised and unsupervised machine-learning models.

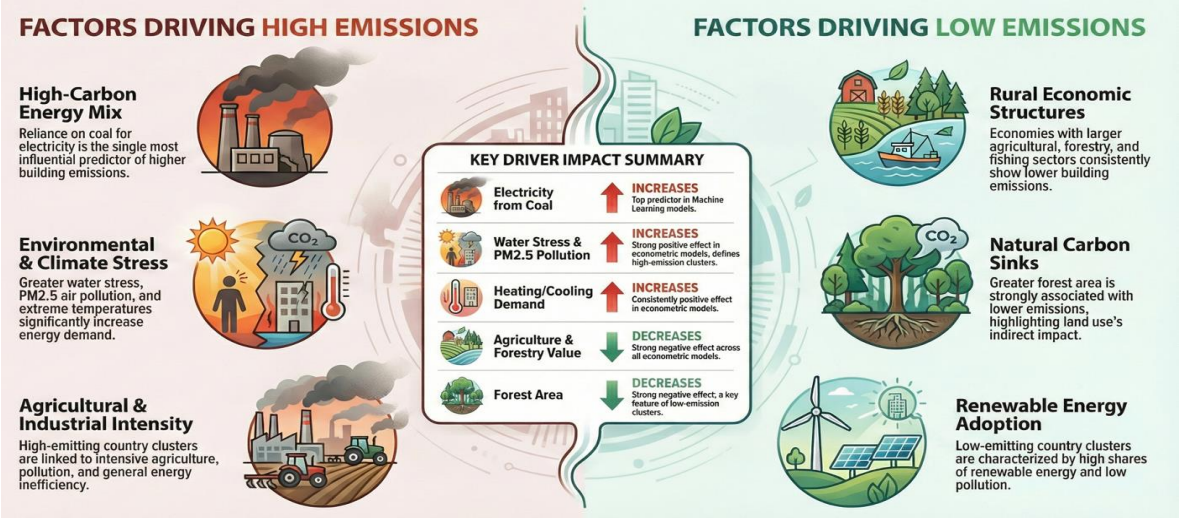

Figure 1. Drivers of High and Low CO_2 Emissions in the EU Building Sector. This figure summarises the main structural, environmental, and energy-system drivers distinguishing high- and low-emission country clusters, highlighting how energy mix, climate stress, land use, and renewables jointly shape building-sector CO_2 emissions across Europe.

3.2. Machine Learning Models and Clustering Analysis

The choice of machine learning algorithms is based on a compromise between interpretability, flexibility, and empirical results. A variety of other regression models: Boosting Regression, Decision Tree Regression, K-Nearest Neighbours Regression, Linear Regression, Neural Networks Regression, Random Forest Regression, Regularized Linear Regression, and Support Vector Regression Machine Regression models are systematically tested and compared based on standard statistical measures of performance: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2).

The performance of models is tested for hold-out test samples, and hyperparameters are tuned for hold-out validations as described in Appendix A. The econometric panel models continue to be the main analytical tool for statistical inference and policy analysis. Machine learning methods are used as a supplementary tool for non-linear modelling of relationships and effects that may not be properly specified by econometric models. In this way, a clustering analysis is performed for further integration of results.

A variety of clustering methods: density-based clustering methods, fuzzy C-means clustering methods, hierarchical clustering methods, model-based clustering methods, and k-means clustering methods are compared based on established cluster validity and separation measures: maximum cluster diameter, minimum separation, Pearson's γ, Dunn index, entropy, and Calinski-Harabasz index. A multi-criteria clustering analysis allows for robust and interpretable results for country groups. Thus, a typology-based analysis of results from econometric models and machine learning models is provided rather than two separate results.

4. Panel Data Modelling Framework for Assessing Emissions Determinants in EU Buildings

To investigate the structural and environmental determinants of carbon dioxide emissions from the building sector (CBE) across 27 European Union countries, we estimate the following baseline specification:

$$CBE_{it} = \alpha_{it} + \beta_1(AFFV)_{it} + \beta_2(FRST)_{it} + \beta_3(WSTR)_{it} + \beta_4(PM25)_{it} + \beta_5(FPI)_{it} + \beta_6(HDD)_{it} + \beta_7(N20P)_{it}$$

Where; $i = 27, t = [2005; 2023]$

The set of variables included in this model describes the economy's structure, capacity, and pressures related to the environment and climate-related energy demand, as outlined in Section 3. Specifically, this set of variables was chosen to represent established relationships between energy and the environment that are applicable to CO_2 emissions from buildings within Member States of the European Union. The empirical analysis focuses on the EU-27, where there is a common climate policy and common frameworks, but also high diversity with regard to climate, geography, and economy. To address this dual nature of data, this analysis utilises a range of panel data methods that are complementary to each other: Fixed Effects, Dynamic Panel GMM, and Weighted Least Squares.

Specifically, we used the following variables to estimate the econometric model (Table 2).

Table 2. Definitions of Variables Used in the Econometric and Machine-Learning Analysis

Acronym	Variables	Definition
CBE	Carbon dioxide (CO_2) emissions from Building (Energy) (Mt CO_2 e)	Measures CO_2 emissions specifically generated by the building sector (heating, cooling, electricity, fuel use).
AFFV	Agriculture, forestry & fishing value added.	Economic output of primary sectors, reflecting productivity and structural economic characteristics.
FRST	Forest area	Total land area covered by forests is an indicator of natural carbon sinks and environmental conservation.
WSTR	Water stress level	Degree of pressure on freshwater resources, representing environmental vulnerability and resource scarcity.
PM25	PM2.5 air pollution	Concentration of airborne particles smaller than 2.5 micrometres is a major indicator of air pollution and environmental degradation.
FPI	Food production index	Composite index measuring changes in agricultural output relative to a base year.
HDD	Heating Degree Days	Heating Degree Days measure how cold a location is over a period of time by calculating the number of degrees that daily temperatures fall below a base threshold (typically 18°C).
N2OP	Nitrous oxide emissions (per capita)	Per capita emissions of N_2O, a potent greenhouse gas mainly from agriculture, waste, and industrial processes.

Note: All variables are sourced from the World Bank and represent structural, environmental, and climatic dimensions relevant to building-sector CO_2 emissions across EU countries. Link: https://data360.worldbank.org/en/search

The empirical approach relies on four different panel data estimators: Random Effects, Fixed Effects, Dynamic Panel GMM, and Weighted Least Squares, which are employed collectively to examine the determinants of carbon dioxide emissions from the building sector (CBE). The selection of the four estimators is based on the key econometric problems that may arise when dealing with cross-country panel data, namely, the problem of heterogeneity, the problem of endogeneity, and the problem of heteroskedasticity. The Fixed Effects approach is employed to control for time-invariant country characteristics that might be correlated with the explanatory variables, while the Dynamic Panel GMM is employed as a robustness check for the possibility of heterogeneity and endogeneity through the use of internal instruments. The Weighted Least Squares approach is employed to correct the heteroskedasticity problem through the use of weights that are based on the variance, which improves efficiency.

The results from the panel data study provide strong empirical evidence for the correlation between structure factors, environment factors, and climate factors and construction-based emissions (CBE) for 27 Member States in the European Union from 2005 to 2023 (Cámara-Aceituno et al., 2025; Petrescu et al., 2023). Regardless of differences in model specifications, including Random Effects, Fixed Effects, Dynamic Panel GMM, and Weighted Least Squares models, there are clear patterns in the results that show a consistently negative coefficient for agriculture, forestry, and fishing value added (AFFV) in most models, suggesting that countries with more primary activity have lower CBE. This is consistent with country differences in building energy consumption patterns in urban settlements in the European Union (Cámara-Aceituno et al., 2025; Doran et al., 2025).

Forest area (FRST) demonstrates a strong negative correlation with CBE, indicating that those land use patterns that support greater natural carbon sequestration are, in general, associated with lower building energy

intensities (Bardulis et al., 2023). By contrast, water stress (WSTR) is positively associated with CBE, suggesting that regions that are water-stressed, such as Southern Europe, are associated with greater energy demands and infrastructure investments in water treatment, transportation, and supply chain infrastructure. Also, PM2.5 concentrations are positively associated with emissions, providing further support that there is an association between fossil fuel-heated buildings and urban area carbon intensity (Doran et al., 2025).

Food Production Index (FPI), in general, demonstrates a positive association with emissions, suggesting an indirect association between energy demand, agri-food supply chain infrastructure, and building energy demand and emissions (Cámara-Aceituno et al., 2025). Nitrous oxide emissions per capita (N2OP), in general, tend to be positively associated with building emissions, except in Weighted Least Squares, where it is negatively associated due to heteroskedasticity. Finally, in the dynamic model, it is observed that there is a modest degree of path dependence in CBE, consistent with the effect of new EU directives regarding energy efficiency in mitigating path dependence in building-related emissions (Petrescu et al., 2023; Roca Reina et al., 2025). In general, these results highlight that there is a combined effect of land use, environmental pressures, and economic structure that influences building-related emissions in the European Union, where a common pattern is observed across models (see Table 3).

Table 3. Regression Estimates Across Four Panel Data Models (GLS, GMM, FE, WLS).

Variables	Random Effects (GLS)	Dynamic Panel (GMM)	Fixed Effects	WLS
Constant	31.401** (13.056)	—	76.376*** (16.576)	12.532*** (3.512)
CBE(-1)	—	0.127 (0.134)	—	—
AFFV	−1.884*** (0.410)	−1.918** (0.801)	−2.366*** (0.430)	−2.160*** (0.363)
FRST	−1.295*** (0.302)	−3.244** (1.548)	−2.564*** (0.457)	−0.473*** (0.032)
WSTR	0.166*** (0.040)	0.230* (0.123)	0.169*** (0.040)	0.424*** (0.033)
PM25	0.506*** (0.099)	0.402** (0.183)	0.374*** (0.104)	0.278*** (0.083)
FPI	0.097*** (0.021)	0.097*** (0.031)	0.111*** (0.021)	−0.047* (0.026)
HDD	0.003*** (0.000)	0.003*** (0.001)	0.003*** (0.000)	0.004*** (0.000)
N2OP	4.317** (2.080)	8.849* (5.128)	4.695** (2.084)	−9.211*** (0.883)

Note: Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. "—" indicates that the variable is not included in the corresponding specification.

A comparison among the four panel data estimators shows that there are significant differences in terms of performance and robustness in estimating the emissions in the European building sector (Bezić et al., 2022). Although there are similarities in terms of the distribution of the dependent variable, there are significant differences in goodness-of-fit tests. In particular, there is a slightly poorer performance in terms of residual variance and information criteria, indicating that it is not the preferred model in terms of characteristics of the considered data. This finding is supported by the Hausman test, favouring the Fixed Effects model, showing that there is a correlation between individual effects and regressors.

Although there is a smaller sum of residuals in Dynamic Panel GMM, due to possible over-identification, GMM tests are questionable, despite the absence of first- and second-order autocorrelation in AR(1) and AR(2) tests (Dincă et al., 2022). Fixed Effects proves its robustness in terms of performance in various tests and statistics, as reflected by lower variance in residuals, larger log-likelihood values, and better information criteria. The high value of R^2 in the LSDV Fixed Effects highlights the significance of country-specific and fixed factors in accounting for diversity in a large group of EU countries, while unit variance reflects a moderately strong level of explanatory power.

Weighted Least Squares estimation also improves efficiency by accounting for heteroscedastic residuals, leading to smaller standard errors and better goodness-of-fit statistics. However, the residuals indicate cross-sectional dependence and non-normal residuals, pointing towards common shocks and common policies in a large number of EU countries—a common phenomenon in integrated regional systems (Petruška et al., 2022; Nichifor et al., 2025). Therefore, while the Fixed Effects and Weighted Least Squares techniques are relatively better, the non-cross-sectional estimation procedure also emphasises the importance of country-specific and interdependent variables that determine the emissions of the member countries of the EU (Bezić et al., 2022; Bonar, 2024).

The empirical results obtained using the quartet of panel data estimators provide a basis for comparing model performance and assessing robustness of structural relationships for CO_2 emissions in the building sector for Member States of the European Union over 2005-2023. Diagnostic tests reveal marked discrepancies among specifications, thus facilitating a sound judgment about the appropriateness of each estimator (see Table 6). Although the Random Effects model shows statistically significant relationships between the explanatory variables and emissions, the Breusch-Pagan test indicates the existence of country-specific effects, thus ruling out the appropriateness of pooled estimation. This result is supported by the Hausman test, which rejects consistency of the Random Effects estimator and recommends the application of Fixed Effects to address correlation between explanatory variables and time-invariant heterogeneity.

The Dynamic Panel GMM specification includes the role of emissions persistence through a lagged dependent variable, thus suggesting a limited path dependence relationship for the emissions of the building sector. However, a failure of the Sargan test indicates over-identification of the instrument and a lack of validity of the GMM test, thus limiting the validity of GMM test findings despite satisfactory tests of AR(1) and AR(2) specifications (Bezić et al., 2022). The Fixed Effects estimator outperforms other specifications along a range of diagnostic tests, including residual variance, log-likelihood, and information criteria, thus suggesting the estimator's effectiveness in adjusting for country-specific heterogeneity for Member States of the European Union.

Evidence of heteroskedasticity and cross-section dependence among specifications suggests a common shock and structural relationships among EU countries. The Weighted Least Squares estimator improves efficiency by adjusting for heteroskedasticity, thus reducing standard errors and improving fit statistics. Although non-normality and cross-sectional dependence remain a concern, these properties are consistent with the integrated nature of the EU's economic and environmental system and do not pose any threat to the substantive interpretation of findings.

Based on findings, the Fixed Effects and Weighted Least Squares estimators can be deemed as the most robust methods for this study, with Dynamic Panel GMM retained for robustness purposes due to limitations of validity of the instrument for GMM test findings. The consistency of findings among specifications lends validity to the role of structural and environmental factors in explaining the level of CO_2 emissions in the building sector for the European Union (Bonar, 2024; Raycheva, 2023; Vagnini et al., 2025). See Table 4.

Table 4. Diagnostic Tests for Panel Data Estimators.

Test	Random Effects	Dynamic Panel (GMM)	Fixed Effects	WLS
Joint significance	$\chi^2(7) = 224.21$, $p \approx 0$	Wald $\chi^2(8) = 195.81$, $p = 0$	$F(7,401) = 34.42$, $p \approx 0$	$F(7,427) = 77.99$, $p \approx 0$
Breusch–Pagan LM	$\chi^2(1) = 3101.46$, $p = 0$	—	—	—
Hausman test	$\chi^2(7) = 17.32$, $p = 0.015$ → FE preferred	—	—	—
AR(1)	—	$z = -1.666$, $p = 0.0956$	—	—
AR(2)	—	$z = 1.230$, $p = 0.2186$	—	—
Sargan over-ID	—	$\chi^2(107) = 293.54$, $p = 0.0000$ → invalid instruments	—	—
Test for fixed vs pooled	—	—	$F(26,401) = 1485.42$, $p = 0$	—
Heteroskedasticity	—	—	$\chi^2(27) = 1.12\times10^8$, $p = 0$	—
Residual normality	$\chi^2(2) = 109.25$, $p \approx 0$	$\chi^2(2) = 556.17$, $p \approx 0$	$\chi^2(2) = 330.32$, $p \approx 0$	$\chi^2(2) = 473.92$, $p \approx 0$
Wooldridge autocorrelation	$F = 0.080$, $p = 0.779$	—	$F = 0.080$, $p = 0.779$	—
Pesaran CD	$z = 5.47$, $p \approx 0$ → cross-sectional dependence	$z = 1.22$, $p = 0.221$ → no dependence	$z = 3.77$, $p \approx 0.00016$	$z = 8.50$, $p \approx 0$
Overidentification/ model diagnostics	—	Instruments invalid	—	—

4.1. Interpretation and Limitations of the Dynamic Panel GMM Results

Nevertheless, the analysis of the GMM outcomes is impeded by some key restrictions. The tests for validity suggest that the Sargan over-identification test is not valid, which affects the internal instruments used for the GMM analysis. However, the outcome is indicative of possible over-identification, which affects the reliability of the GMM analysis on the causal relationship among the variables studied. In addition, the problem has been identified among previous studies that have used the dynamic panel GMM estimators within the context of the environment (Iqbal et al., 2025; Suproń & Myszczyszyn, 2024). As such, the Dynamic Panel GMM approach should focus on offering additional insights and act as a robustness check for other models (Nigmatullaeva et al., 2025).

In contrast, the Fixed Effects and Weighted Least Squares models have relatively high levels of diagnostic accuracy and robustness tests. As such, these models are the most appropriate for the analysis of the average relationship among the structural, environmental, and climate variables and the CO_2 emissions from the building sectors among the Member States of the EU.

4.2. Diagnostics

However, it should not be surprising that there is heteroskedasticity and cross-section dependence in a panel of countries such as the EU-27, where there are large variations between countries in terms of economic structure, climate patterns, energy infrastructure, and susceptibility to common shocks through EU policies. In a highly integrated economic and institutional system such as the EU, cross-section dependence can be primarily caused by common shocks, spillovers, and common policies such as EU common climate and energy policies and common energy markets (Pesaran, 2021; Rietig, 2021; Okunevičiūtė Neverauskienė et al., 2025).

Heteroskedasticity refers to the heterogeneity of country emissions with varying levels of economic size, institutional capabilities, and structural attributes, a phenomenon that has been widely found in cross-country environmental panel data (Usman & Jahanger, 2021; Awad & Warsame, 2022). From a methodological perspective, these concerns were directly dealt with within the empirical approach. First, the Fixed Effects estimator accounted for time-invariant country heterogeneity, which helped correct for bias due to unknown structural disparities among Member States, even when there are disparate emission paths (Usman & Jahanger, 2021).

Second, the Weighted Least Squares (WLS) method was used specifically for dealing with heteroskedasticity, which improved the efficiency and robustness of conventional standard error estimates, especially when the variance is systematically disparate across cross-sectional entities (Awad & Warsame, 2022). Third, the joint sign and significance of the coefficients across various estimators helped verify that the key findings were robust and not dependent on a particular modelling assumption (Pesaran, 2021).

Turning to interpretation, the presence of detected heteroskedasticity and cross-section dependence does not impinge upon substantive interpretation but instead captures the integrated and interdependent nature of the economic, energy, and policy structures of the European Union (Rietig, 2021; Okunevičiūtė Neverauskienė et al., 2025). Thus, the coefficient estimates must be viewed as averages for interdependent Member States, and not as isolated country-level causal relationships. In this regard, the study's aim of determining structural, environmental, and climatic factors affecting building sector emissions within a common policy and economic framework is well captured.

In general, the findings remain informative and valid, and the use of multiple estimators ensures that inference is not based on a specific modelling condition.

5. Evaluating Machine-Learning Models: KNN as the Benchmark Predictor

Analysis of normalised performance measures clearly shows that the k-Nearest Neighbours algorithm is the most efficient among the set of models analysed (Uddin et al., 2022). This efficacy is evidenced by its dominance on the key predictive accuracy measures. Specifically, it registers the lowest values on normalised measures of mean squared error, Root Mean Squared Error, Mean Absolute Error, and Mean Absolute

Percentage Error. Notably, these measures indicate KNN's efforts to minimise average and absolute prediction errors relative to other models.

Moreover, the MAPE measure, normalised to zero, clearly shows that it is the most precise predictor of percentage forecasts, and this consideration matters significantly in the context of percentage error measurement between predicted and actual observations. Conversely, it yields the highest R^2 value with a measure normalised to 1, indicating it explains almost all changes in the sample variable with respect to the dependent variable (Akakpo et al., 2024).

The Decision Tree algorithm ranks second-best, with extremely low error rates and a remarkably high R^2 value. The Random Forest algorithm remains a close second, with an excellent bias-variance trade-off, but fails to outperform KNN and Decision Tree on any specific criterion (Babbar et al., 2023). Linear Regression and its regularised forms achieve moderately good predictive performance, indicating sensitivity to nonlinearities and high-order interactions within the data. Boosting algorithms offer advantages over traditional linear approaches but cannot match the accuracy of tree- or instance-based models.

Support Vector Machine and Neural Network models rank relatively low on all criteria, with the latter performing poorly across almost all fronts, hinting at issues of either excessive specialisation and ill-regularisation or inappropriateness within its applicational domain. These observations qualitatively align with previous comparative analyses on optimising KNN instances, with highly flexible adaptability to nonlinear similarities being an added advantage for better accuracy across various learnable platforms and datasets, as seen in previous works in similar domains and settings (Ozturk Kiyak et al., 2023; Abualhaj et al., 2024).

Net result, KNN stands out as the most consistent and efficient model among the ones compared. The performance achieved with KNN suggests that neighbourhood similarity relationships within the feature space contain considerable predictive information about the target attribute, making it well-suited for this particular data set and task (Uddin et al. 2022; Akakpo et al. 2024). See Table 5.

Table 5. Performance Metrics for Machine-Learning Regression Models

Model	MSE	RMSE	MAE	MAPE	R^2
Boosting	0.709	0.823	0.588	0.339	0.539
Decision Tree	0.041	0.128	0.131	0.053	0.957
KNN	0.000	0.000	0.000	0.000	1.000
Linear Regression	0.372	0.562	0.677	0.690	0.608
Neural Network	0.920	0.954	1.000	1.000	0.000
Random Forest	0.053	0.154	0.204	0.175	0.890
Regularized Linear	0.501	0.671	0.732	0.503	0.640
Support Vector Machine	1.000	1.000	0.730	0.215	0.502

In particular, the following variables were used to estimate the value of BCE. See Table 6.

Table 6. Variables Used for Estimating Building-Sector CO_2 Emissions (CBE).

Acronym	Variable	Short Explanation
ASND	Adjusted savings: natural resources depletion	Measures the monetary value of natural resource depletion (minerals, forests, fossil fuels), reflecting environmental sustainability pressures.
AGRL	Agricultural land	Share of total land area used for agriculture; indicates land-use patterns and economic structure.
AFFV	Agriculture, forestry & fishing value added.	Economic output of primary sectors is associated with rural structures and lower energy-intensive building demand.
AFWW	Annual freshwater withdrawals	Total freshwater used annually across sectors, indicating pressure on water resources.
CO2P	CO_2 emissions (per capita)	Per capita emissions from all sectors; proxy for the general environmental footprint and energy intensity.
CDD	Cooling Degree Days	Measures heat intensity by summing temperatures above a base (typically 18°C); higher values imply greater cooling demand.

Table 6 (Cont.). Variables Used for Estimating Building-Sector CO_2 Emissions (CBE).

Acronym	Variable	Short Explanation
ECL	Electricity from coal	Share of electricity generated from coal signals reliance on high-carbon energy.
EIMP	Energy imports, net	The difference between imported and exported energy indicates energy dependency and vulnerability.
ENIN	Energy intensity	Energy use per unit of GDP signals the efficiency of economic production.
ENUP	Energy use (per capita)	Total energy consumption divided by population reflects lifestyle, climate, and industrial structure.
FPI	Food Production Index	Indicates agricultural output relative to a base year; captures agricultural productivity trends.
FRST	Forest area	Total land area covered by forests reflects carbon-sink capacity and environmental conservation.
FOSC	Fossil fuel consumption	Share of energy from fossil fuels; proxy for carbon intensity of the energy system.
HDD	Heating Degree Days	Measures cold intensity by summing temperatures below a base threshold; higher values imply greater heating demand.
WSTR	Water stress level	Degree of pressure on freshwater availability relative to demand; an indicator of environmental vulnerability.
CH4P	Methane emissions (per capita)	Per capita methane emissions; reflects agricultural and waste-sector contributions to GHGs.
N2OP	Nitrous oxide emissions (per capita)	Per capita N_2O emissions, mainly from agriculture and industry; a potent greenhouse gas indicator.
PM25	PM2.5 air pollution	Concentration of fine particulate matter (<2.5 μm), associated with fossil fuel combustion and urban pollution.
RELE	Renewable electricity output	Share of electricity generated from renewable sources; an indicator of clean-energy transition.
RENC	Renewable energy consumption	Share of final energy derived from renewables reflects national commitment to decarbonisation.
TCL	Tree cover loss	Annual forest cover loss signals deforestation and reduced carbon-sink capacity.
CBE	Carbon dioxide emissions from buildings (Mt CO_2e)	CO_2 emissions produced by building energy use (heating, cooling, electricity, fuels).

The results obtained via the K-Nearest Neighbours algorithm provide useful information on the importance of large sets of variables employing various environmental, energy, and socioeconomic factors for modelling and understanding carbon dioxide emissions from the building sector (CBE) (Yancey et al., 2021). The average values of dropout loss serve as an indication relating to variable importance within a prediction task, and they are associated with larger sensitivity due to removal for prediction modelling or loss within respective variables (Hasler and Tillé, 2016).

The importance of these variables obtained via modelling shows a hierarchical importance pattern reflecting theoretical considerations for energy demand and emission drivers as per the theoretical perspective on energy and emission drivers in Europe. The variable indicating electricity production via coal (ECL) becomes highly influential, reflecting considerations on fossil fuel-based energy and emission production drivers on emission profiles and energy production drivers surveyed and documented via recent research reviews (Li et al., 2021). Even though there is a decrease in coal usage in several economies within Europe, economies dependent on coal usage demonstrate significantly higher CBE compared to economies with no dependence on these resources.

Water stress factors appearing as variable WSTR and value-added agriculture AFFV demonstrate relevance, reflecting considerations on energy consumption drivers under similar conditions and pressures appearing due to variable determinants among regions and economies—factors usually reflecting prevailing eco-pressures and drivers among regions and economies dependent on and separated due to controlled agriculture and ecological variability factors and determinants among regions and economies surveyed and documented via research reviews.

Cooling degree-days CDD appear next, reflecting considerations on energy consumption drivers and factors acquiring relevance and causality due to rising temperature, eco-pressures and variability among regions and economies surveyed and documented via research reviews. Variables reflecting environmental factors appear highly influential, considering factors CH4P, PM25, and TCL variable influence on CBE within regions and economies surveyed and documented via research reviews. Energy variables, demonstrating factors ENIN, EIMP, and FOSC variables among regions and economies, act as influential factors on CBE among regions and surveyed economies, as documented via research reviews.

On the other hand, variables like consumption of renewable energy (RENC), per-capita energy use (ENUP), and forest area (FRST) have relatively low dropout loss values, implying that, while these variables are useful in modelling, they have relatively less effect as a marginal contribution after taking into consideration other structural variables. All in all, KNN outputs demonstrate a highly intricate relationship existing among various climatic factors, energy structures, pollutant levels, and various land use factors, suggesting that greenhouse gas emissions related to buildings are circumscribed within larger environmental and economic frameworks (Yancey et al. 2021; Li et al. 2021; Mohammed et al. 2017). See Table 7.

Table 7. Mean Dropout Loss Ranking of Variables in the Machine-Learning Model

Variables	Mean dropout loss	Variables	Mean dropout loss
ECL	14.433	HDD	3.337
WSTR	7.562	ASND	3.175
AFFV	6.193	N2OP	2.933
CDD	6.044	FOSC	2.823
AFWW	5.797	RELE	2.811
CH4P	5.415	CO2P	2.705
PM25	5.408	FRST	2.652
TCL	4.598	AGRL	2.609
ENIN	4.590	RENC	2.556
EIMP	4.514	ENUP	2.440
FPI	3.449		

The table records five prediction scenarios created by the model, shown as comparisons of the predicted values on carbon emissions from buildings with 21.114 as the baseline and illustrating changes made in each explanatory variable that affect the prediction. It shows how a set of variables on environmental factors, energy, and land use affects the prediction either positively or negatively against the 21.114 baseline (Guo & Luo, 2024; Hsu et al., 2022).

Case 1 presents an expected result much below the benchmark. The major suppressant forces include large negative factors of water stress (WSTR), PM2.5 air pollution, renewable energy production (RELE), and tree cover loss (TCL). Suppressive forces outweigh promotional factors of cooling degree days (CDD) and fossil fuel consumption (FOSC). The trend here shows a country with strong environmental factors but moderate pressures associated with climate variables (Dai et al., 2024).

Case 2 shows a comparable downward pattern, again because of negative factors related to water stress and RELE, but with a stronger positive influence from AFFV and CDD. The negative influence of HDD represents a warmer and drier climate with low heating requirements. The strong positive role of forest area (FRST) represents an attribute that decreases GHG emissions indirectly (Yao et al., 2024).

Case 3 presents an enormous increase in projected emissions, surpassing the level of the base case significantly. The overriding factors here include large positive impacts of AFFV and FPI, together with high levels of water stress and PM2.5 concentrations. The large negative roles of ENIN and EIMP suggest inefficiencies and dependence on external energy sources, but these alone are not strong enough to offset the overriding factors. All these factors will be reflected again by the persistence seen in energy and agriculture pressures within Europe, as given in Kayakuş et al. (2023).

Also predicted under Case 4 are emissions beyond the base, with drivers mainly due to AFFV, forest area, water stress, and specifically high levels of RENC. However, the large negative weights for PM25 and TCL

offset these effects. Emissions distribution fits previous research about discrepancies regarding energy consumption/abatement factors for urban vs. rural regions (Hsu et al., 2022).

Case 5 shows the least amount of predicted emissions, significantly lower than the baseline, due to very low values for CDD, AFFV, ENIN, and RELE. The pattern reflects a low-emission scenario with low energy demands and a supportive environment, consistent with forecasts made by optimised machine learning algorithms incorporating factors of environmental and land use changes (Dai et al. 2024; Yao et al. 2024). See Table 8.

Table 8. Variable-Specific Contributions to Predicted Building-Sector CO_2 Emissions Across Five Cases.

Case	Predicted	Base	ASND	AGRL	AFFV	AFWW	CO2P	CDD	ECL	EIMP	ENIN
1	9.284	21.114	0.340	0.080	0.164	-1.506	0.332	2.684	-2.071	0.119	-0.073
2	8.434	21.114	0.160	0.521	2.622	-1.400	-0.011	2.624	-2.859	0.144	-0.327
3	26.966	21.114	-0.052	2.103	5.480	-0.990	-0.327	0.966	-2.816	-3.480	-5.989
4	23.848	21.114	-0.066	0.929	5.570	-3.714	-0.567	-0.795	-3.351	0.160	-3.693
5	1.532	21.114	-0.698	-0.022	-5.075	-0.918	-0.554	-7.622	2.014	-0.449	-5.003

Case	ENUP	FPI	FRST	FOSC	HDD	WSTR	CH4P	N2OP	PM25	RELE	RENC	TCL
1	0.081	0.312	0.185	1.382	1.486	-7.827	-0.111	0.634	-0.273	-5.852	-0.470	-1.448
2	0.232	0.049	1.289	-0.063	-4.362	-4.189	-0.077	0.407	-0.049	-5.510	-0.932	-0.950
3	0.619	7.211	-0.404	-0.081	-0.085	5.239	-0.033	0.413	0.686	1.054	1.040	-4.702
4	1.208	0.397	1.042	0.005	1.067	3.863	2.270	-0.062	-3.907	0.000	7.232	-4.856
5	-1.089	-1.033	0.000	-0.006	-0.577	0.767	-0.036	-1.361×10^{-4}	0.736	0.000	-0.004	-0.015

Figure 2 below shows two different complementary graphs that display the functionality of a K-Nearest Neighbours Regression model. The graph on the left compares and contrasts predicted and actual values. From the graph, it can be seen that all points lie almost on the red line that represents perfect prediction. It shows that the K-Nearest Neighbours Regression model perfectly captures the relationship and produces outputs that are very close approximations of actual values (Mailagaha Kumbure and Luukka, 2022). A slight deviation for larger values can be seen.

The plot on the right side investigates the model performance with varying values of k, the number of neighbours. Both the training and validation curves of the mean squared error (MSE) have a typical U-shape due to bias–variance trade-off, which is common for KNN models (Nader et al., 2022). At a very small value of k, there will be a possibility of overfitting, as noticed from the variability and the large gap between the train and validation error values. As k increases, there will be a decrease in error on the validation set, reaching a trough at about k = 4, marked with a red dot. It can be noticed that using neighbours equal to four will result in optimal generalisation on the given data set (Levada et al., 2024). After that, there will be an increase in error, meaning that after a certain value, over-smoothing occurs, and there will be no chance to identify relevant patterns.

Overall, these two graphs indicate that the KNN algorithm performs predictively well and that there is an optimal value for k. Both graphs confirm that KNN is an appropriate algorithm for the given prediction task (Mailagaha Kumbure and Luukka, 2022; Nader et al., 2022; Levada et al., 2024). See Figure 2.

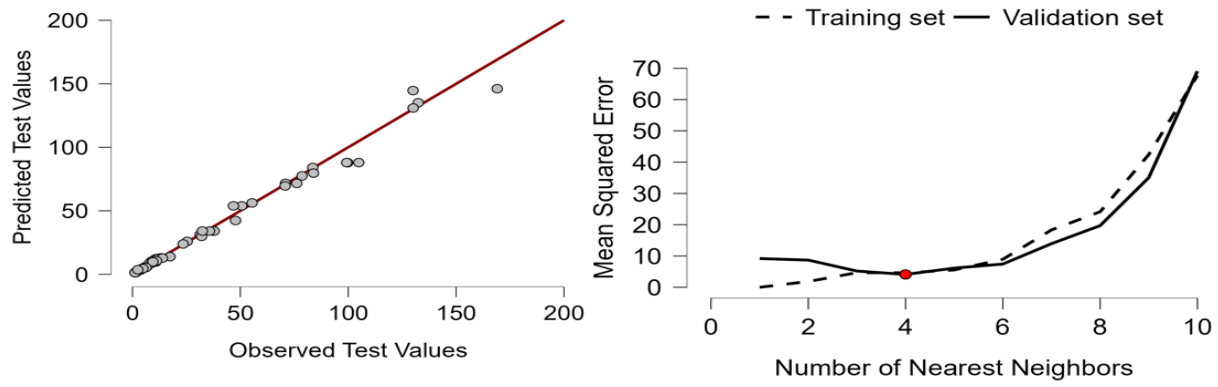

Figure 2. Performance of the K-Nearest Neighbours Model: Prediction Accuracy and Optimal Neighbour Selection

6. Evaluating Clustering Algorithms and Selecting the Optimal Hierarchical Model

To identify which clustering algorithm performs best, it is necessary to refer back to these normalised values as indicative of clustering performance and not combine them into a single value. Each value highlights a different characteristic of the structure represented by these clusters. It therefore becomes necessary that the best algorithm would be that which performs well on as many relevant factors as possible, as identified in Ikotun et al. (2025).

It can be seen that the Hierarchical algorithm boasts the lowest maximum diameter at 0.000 and represents the greatest level of maximum diameter. It also boasts one of the greatest measures of minimum separation at 1.000, representing highly separated clusters. It also boasts the greatest value for Pearson's γ at 1.000, representing highly internally correlated and highly structured clusters as per Arbelaitz et al. (2014).

Its Dunn value at 1.000 represents excellent performance, and although there is room for improvement at 0.186, its entropy value is moderate and acceptable. Its CH value at 0.920 represents highly advantageous cluster validity. It falls slightly short, as it represents the second-best algorithm on this value. Almost all structural factors appear reliable and consistent as per Ikotun et al. (2025).

On the other hand, fuzzy C-Means clustering shows very low ranks on some criteria, but these ranks relate to poor performance because they represent normalised distance measures compared with the best algorithm. Its minimum separation and Dunn value are very low, indicating that it poorly delineates regions among clusters, with its Calinski-Harabasz index also being the lowest among all.

The density-based clustering algorithm shows mediocre performance but very low entropy and cluster validity indices. Model-based, k-means, and random forest clustering algorithms show varying levels of performance because they display good ranks on some criteria but rank poorly on others, indicating that they might be inconsistent and algorithm- or structure-dependent, as argued by Zangana and Abdulazeez (2023).

Based on the sum of the strongest compactness, separation, and structural validity, hierarchical clustering clearly ranks as the best-performing method. Its trade-off between tight clustering and clear partitioning makes it the most trustworthy among other methods, especially if there are hierarchical relationships within the data sets that provide valuable interpretations (Arbelaitz et al. 2014; Ikotun et al. 2025). See Table 9.

Table 9. Normalised Cluster-Quality Metrics Across Six Clustering Algorithms.

Metric	Density Based	Fuzzy C-Means	Hierarchical	Model Based	k-Means	Random Forest
Maximum diameter	0.352	0.825	0.000	1.000	0.444	0.626
Minimum separation	0.045	0.000	1.000	1.000	1.000	1.000
Pearson's γ	0.376	0.000	1.000	0.318	0.947	0.741
Dunn index	0.065	0.000	1.000	0.751	0.874	0.833
Entropy	1.000	0.290	0.186	0.000	0.265	0.265
Calinski–Harabasz index	0.915	0.000	0.920	0.370	1.000	0.708

The cluster information allows for a detailed representation of the structure and quality of the solution with ten clusters. The size of these clusters varies from relatively large groups, as seen in Cluster 1 with 80 observations, to relatively smaller groups with 16 observations. The difference in cluster size suggests that there are broader and more specific micro-structures within the dataset, as suggested by Sasmita et al. (2023).

The proportion of explained within-cluster heterogeneity varies as per cluster size; relatively larger clusters, as seen in Cluster 1, explain more variability (0.391), while relatively smaller clusters, as seen in clusters 8, 9, and 10, explain relatively low proportions. Nonetheless, it is normal because smaller clusters explain more homogeneous micro-structures. The within-sum of squares shows the compactness and homogeneity within these clusters. It appears that clusters 7, 8, 9, and 5 have relatively low within-cluster variability, suggesting relatively tight and more homogeneous groups. Conversely, relatively larger clusters 1, 2, and 6 have relatively higher within-cluster variability, as seen with larger and more homogenous datasets (Park et al., 2025).

Silhouettes offer critical insights into cluster separation and cohesion. A higher value represents a proper fit and assignment within these clusters. Higher values within these clusters suggest relatively better fitments.

Notable here are clusters 8 and 9 with relatively high values of 0.897 and 0.828, respectively. It emphasizes highly cohesive and relatively distant clusters. Also relatively good is Cluster 5 with a value of 0.660.

Conversely, relatively less cohesive/less fit within these clusters is Cluster 1 with a value of 0.321. It may be due to relatively larger and more diverse cluster-size character. However, average silhouettes convey an indication that while relatively more stable clusters exist here, some might be relatively less distinct or have micro-structure warranting additional refining steps required here (Lenssen & Schubert, 2024). Lastly, relative size and proportioning as per total variability, as seen with within-between cluster sum of squares 5404.72 and total variability 8404, it can be determined here that relatively large proportions are explained here.

It thus reinforces relatively stable representation at the ten cluster solution level and emphasizes relatively more pronounced variability within these datasets here as well and here at different levels suggested by Park et al. (2025), and also suggested by Sasmita et al. (2023). See Table 10.

Table 10. Summary Characteristics of the Ten Clusters Identified in the Building-Emissions Dataset

Cluster Information										
Cluster	1	2	3	4	5	6	7	8	9	10
Size	80	48	48	32	16	63	32	16	16	32
Explained proportion within-cluster heterogeneity	0.391	0.122	0.119	0.077	0.035	0.115	0.062	0.017	0.028	0.033
Within sum of squares	1.172	365.175	357.240	232.330	104.413	345.248	185.997	52.144	84.454	100.230
Silhouette score	0.321	0.457	0.487	0.527	0.660	0.481	0.610	0.897	0.828	0.571

The result obtained from clustering provides a comprehensive explanation of the differences among sets of observations regarding environmental, economic, and energy factors, as well as respective levels of CBE. All clusters carry a unique set of predefined variable values reflecting clearly identifiable drivers for high/low emission factors (Liu et al. 2024). Cluster 4 shows the maximum value for CBE, associated with large positives for AFFV, AFWW, CO2P, and N2OP.

The cluster appears to identify economies with more agricultural production and higher pollution, side by side with less energy efficiency, reflected by a correspondingly low value for ENIN. A high level of emissions might be, at least partially, caused by an energy-draining economic structure (Li et al. 2023; Song et al. 2024). By contrast, CBE values in Clusters 8 and 9 are shown to be the lowest. C8 depicts very large cooling degree days (CDD = 4.341), indicating a climate with large cooling demands, but exhibits highly negative values for emission-linked variables like CO2P and CH4P. It can be gathered that these two factors are associated with an efficient energy sector or focus on renewable energy sources, thus offsetting emission effects despite high demands due to climatic conditions. C9 also depicts low CBE due to highly negative fossil fuel consumption and PM2.5, but with an unusually large value for electricity from coal (ECL = 3.533).

It may be a situation with transitional economies, where there are shares of coal energy, but relatively low compared to expectantly larger contributions towards building emission sources due to sectoral electrification and very rapid advancements in heating technologies (Kosowski, 2024). Cluster 1 presents moderately low emission levels, yet with a remarkable consumption level for RENC = 1.019 and large water stress WSTR. The seeming contradiction here points towards an environmentally constrained setting that offsets the pressure from climate change with an increase in RENEWABLE ENERGIES (Doryń & Wawrzyniak, 2024). Cluster 3 shows almost average emission levels, with high energy imports EIMP = 1.841. Perhaps an energy structure deficit drives energy consumption. Cluster 5 reveals low CBE levels with large agricultural land and AFFV, which could indicate that rural or agrarian regions are associated with low MNB because of low populations and, thus, less heating and cooling demands.

The ten clusters indicate considerable diversity with regard to the emissions - some factors associated with agriculture, pollution, and inefficient energy use, while low emissions represented factors such as the adoption of renewable energy, low pollution, and some specific climatic conditions. As pointed out by Li et al. (2023), Kosowski (2024), and Doryń & Wawrzyniak (2024), there exist clusters associated with high and low emissions. See Table 11.

Table 11. Standardised Cluster Centroids for Environmental, Structural, and Energy Variables

	CBE	ASND	AGRL	AFFV	AFWW	CO2P	CDD	ECL	EIMP	ENIN	RENC
Cluster 1	0.065	-0.693	-1.254	-0.434	-0.553	-0.819	-0.217	-0.389	-0.837	-0.087	1.019
Cluster 2	-0.958	0.393	0.294	-0.599	1.811	-0.430	-0.439	0.147	-0.274	0.374	-8.592×10^{-4}
Cluster 3	0.663	0.026	0.202	0.418	0.002	0.041	-0.004	0.288	1.841	-0.927	-0.140
Cluster 4	1.626	-0.484	-0.104	1.778	-0.474	0.743	-0.323	-0.881	0.059	-0.827	-0.447
Cluster 5	-0.846	-0.264	1.457	2.031	-0.524	-0.985	1.021	0.050	0.565	-2.306	-0.554
Cluster 6	0.140	0.098	0.289	-0.489	0.236	1.672	-0.254	-0.326	0.028	0.908	4.024×10^{-4}
Cluster 7	0.325	2.869	0.895	0.446	0.015	0.106	-0.295	-0.025	-0.061	-0.178	-0.531
Cluster 8	-0.992	-0.762	1.375	-0.448	-0.394	-1.143	4.341	0.422	-0.038	1.131	-0.526
Cluster 9	-1.524	-0.650	0.452	-0.599	-0.627	-0.643	0.005	3.533	-1.102	1.915	-0.620
Cluster 10	-0.266	-0.637	-0.610	-0.396	-0.570	-0.123	-0.358	-0.135	-0.025	-0.106	-0.509

	ENUP	FPI	FRST	FOSC	HDD	WSTR	CH4P	N2OP	PM25	RELE	TCL
Cluster 1	0.383	0.381	-1.151	-0.348	1.271	1.466	0.711	-0.793	0.995	1.439	-0.761
Cluster 2	0.056	0.544	-0.322	0.017	-0.454	-0.285	-0.384	-0.297	-0.655	-0.718	1.438
Cluster 3	1.179	-0.260	0.541	-0.206	-0.156	0.073	-0.300	1.284	-0.999	-0.560	0.964
Cluster 4	-0.367	-1.095	0.182	0.017	-0.252	-0.359	-0.452	0.888	0.688	0.538	-0.899
Cluster 5	-1.182	-0.236	-0.028	0.217	-1.311	0.173	0.639	-0.705	0.759	0.618	0.177
Cluster 6	-0.752	-0.694	0.589	0.467	-0.139	-1.368	-0.720	-0.093	0.021	-0.191	0.539
Cluster 7	0.093	0.043	0.613	-0.398	-1.208	-0.409	-0.201	0.269	-1.024	-0.857	-0.319
Cluster 8	-1.636	-0.338	1.373	-0.023	-1.595	-0.156	3.267	-1.083	-0.468	-1.009	-0.669
Cluster 9	-0.940	2.828	0.807	0.058	-0.083	0.040	0.009	-0.748	0.944	-0.952	-0.890
Cluster 10	0.825	-0.085	-0.479	0.490	0.967	0.089	-0.639	0.795	-0.328	-0.314	-0.853

Figure A above illustrates the model selection process adopted to identify an optimal number of clusters based on three methods: AIC, BIC, and within-cluster sum of squares (WSS). All three methods display a downward pattern as the optimal number of clusters increases from two to ten, indicating that as more groups are formed, accuracy and fit improve as more homogeneous and smaller groups are created (Zhang & Lin, 2021). However, it should be noted that while it becomes more optimal with regard to its reduction on WSS, there should also be consideration for complexity.

Bayesian Information Criterion (BIC), which is illustrated by the dashed line, displays considerable complexity costs for excessive generalizability. As indicated by the red dot, it reached its lowest instance at an optimum cluster solution of ten, thus adhering to optimal accuracy and simplicity at the same time (Rossbroich et al., 2022). Although it continuously decreases for all instances, there is also considerable levelling on BIC at an eighth cluster solution, yet achieved its optimum at ten, thus suggesting there are meaningful structural breaks within the observation that correspond better with a more detailed clustering solution. Panel B illustrates a graphical cluster representation within a lower-dimensional representation, most likely achieved via PCA or tSNE. As can be seen, it becomes easy to analyse and differentiate via colour coding, with points illustrating relatively separate and clearly definable sections without considerable intersection (Choi et al., 2022).

Clusters 1, 4, and 6 display considerable cohesiveness and well-defined points without visible intersection, yet clusters 3 and 5 display moderate intersection, yet still separate and distinctly apart from adjacent sections. As a visible spatial structure, it reinforces and upholds previous statements made on figure A that an optimal solution at ten clusters effectively identifies and characterises heterogeneity within observations. Subsequent labelling made more interpretable on specific observations, per se, enhances determination and understanding as it relates various observation points and respective cluster structure attributions.

As it appears from these two images, there seem to be considerable and valid observations on choosing an optimal cluster at ten. Both measures reflecting optimal BIC and widely separate sections indicate that an optimal cluster solution at ten displays considerable and substantial observations on effectively characterising multidimensional implications on specific observation points as posited on previous assertions specified within various instances (Rossbroich et al., 2022; Zhang & Lin, 2021). See Figure 3.

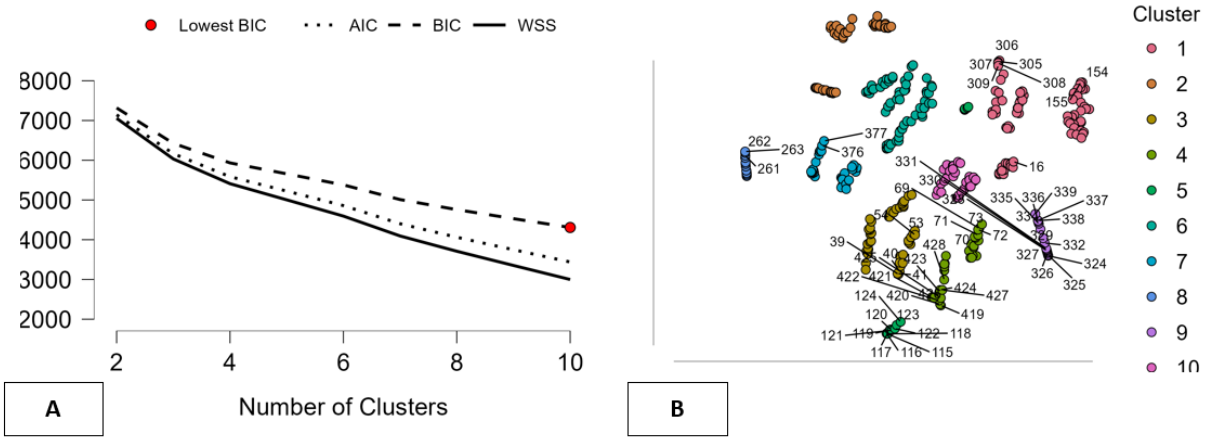

Figure 3. Cluster Selection and Visualisation of the Ten-Cluster Solution

7. Integrated Insights from Econometric, Machine-Learning, and Clustering Evidence

The integrated econometric, machine learning, and clustering analysis demonstrates that a comprehensive and holistic understanding can be achieved regarding the determinants of CO_2 emissions from the building sector (CBE) across the 27 European Union Member States over the period 2005–2023. The triangulation of these methodological approaches highlights that the multidimensional nature of emissions cannot be adequately captured through a single analytical lens, a limitation similarly noted by Biancalani et al. (2024). Instead, the combination of statistical inference, predictive modelling, and structural classification enables a more nuanced interpretation of the drivers underlying emissions patterns.

Econometric findings reveal several strong and statistically significant relationships among explanatory variables. Most notably, an inverse association is observed between agriculture, forestry, and fishing value added (AFFV), forest area (FRST), and building-sector emissions. This suggests that countries with more developed rural economies and greater forest coverage tend to exhibit lower CBE levels. These results reinforce the importance of land-use structures in shaping energy consumption patterns, settlement density, and heating demand, all of which influence national emission trajectories (Vagnini et al., 2025). Conversely, variables such as water stress (WSTR), PM2.5 concentrations, heating and cooling degree-days, and N_2O emission rates display strong positive relationships with CBE. Together, these findings underscore the interconnected nature of climatic stressors, environmental degradation, and energy demand in national building stocks, indicating that both cold-climate heating requirements and warm-climate cooling demands contribute substantially to emissions intensity.

From a methodological standpoint, Fixed Effects and Weighted Least Squares (WLS) estimators emerge as the most reliable econometric specifications. Their robustness in the presence of heteroscedasticity and cross-sectional dependence makes them particularly suited to small-N panel structures characterised by pronounced institutional and structural heterogeneity across European states (Biancalani et al., 2024; Vagnini et al., 2025). In contrast, Dynamic Panel GMM models produce invalid instruments, pointing to deeper methodological limitations in causal modelling of complex environmental systems. The exclusion of instrumental-variable approaches is therefore theoretically justified: within globally interconnected environmental systems marked by feedback loops and interdependence, the existence of valid exogenous instruments detached from these processes becomes conceptually implausible.

Machine learning results complement the econometric analysis by capturing nonlinearities and interaction effects that traditional models may overlook. The superior performance of the K-Nearest Neighbours algorithm suggests that emissions patterns within the EU exhibit strong similarity structures, whereby countries with comparable socioeconomic and environmental attributes tend to display analogous emissions profiles (Giannelos et al., 2023; Giannelos et al., 2024). Variable importance measures further emphasise the role of coal-based electricity generation, water scarcity, agricultural activity, particulate pollution, and climatic conditions, reinforcing the view that CBE outcomes are shaped by overlapping economic, environmental, and energy-system determinants (Tudor et al., 2025).

Clustering analysis adds an interpretative dimension by identifying ten distinct emissions profiles among EU Member States. High-emission clusters are generally characterised by intensive agricultural land use, elevated pollutant levels, and comparatively low energy efficiency, whereas low-emission clusters tend to combine higher renewable energy shares, cleaner environmental conditions, and more favourable climatic contexts. These groupings reveal that decarbonisation pathways across the EU are not uniform but instead reflect structurally differentiated national configurations (Vagnini et al., 2025).

Taken together, the results indicate that greenhouse gas emissions from the EU building sector are driven by complex interactions among economic structures, climatic pressures, environmental conditions, and energy-system characteristics. The three-pillar methodological framework demonstrates that no single modelling paradigm can fully capture these dynamics. Instead, the integration of econometric inference, machine learning prediction, and clustering-based structural analysis provides a more realistic and policy-relevant representation of emissions behaviour within the European Union (Giannelos et al., 2023; Biancalani et al., 2024; Tudor et al., 2025). The empirical results supporting these conclusions are summarised in Table 12.

Table 12. Integrated Interpretation of Key Drivers of Building-Sector CO_2 Emissions Across Methods

Variables	Econometric Effect	ML Importance	Cluster Evidence	Interpretation
AFFV	Strong negative across all models	High dropout loss	Low-emission clusters show high AFFV	Rural economies reduce building energy demand
FRST	Strong negative overall	Moderate	Low-emission clusters have large forest cover	Natural sinks & sustainability alignment
WSTR	Strong positive	Very high	High-emission clusters with severe stress	Environmental vulnerability increases energy use
PM25	Strong positive	High	High pollution clusters	Pollution-energy nexus via fossil fuels
FPI	Mostly positive	Moderate	High-emission clusters	Agro-industrial energy demand
N2OP	Positive in most models	Moderate	High in high-emission groups	Agricultural & industrial emissions
HDD	Strong positive	Top predictors	Climatic clusters differ	Temperature extremes raise thermal demand
Energy System Vars	—	Highest importance	Coal-heavy clusters = high emissions	Electricity mix crucial for decarbonisation

8. Strategic Policy Priorities Informed by Structural, Climatic, and Energy-System Dynamics

The present research moves the agenda forward on decarbonisation in the sector. It appears that CO_2 emissions caused by the building sector have roots in structural factors and interact with pressure factors, climate factors, and energy factors (Cazcarro et al., 2022). Due to multiple factors influencing and being influenced by related factors, it requires an integrative and differentiated approach at the level of Member States.

The continued negative effects of agriculture, forestry, and fishing value added, and forest area absorbed make it clear that land use and structure have a significant role as a conditioning factor for GHG emissions. A strategy that preserves forests and uses sustainable land management practices exerts an indirect contribution to GHG emission reductions via improving natural GHG absorption and relieving pressure on energy-guzzling urbanisation drivers (Fotiou et al., 2024). By extension, rural development plans can be an auxiliary factor for

GHG emission reduction by abating energy-guzzling building stocks, indicating that EU cohesion and rural policy may offer additional benefits on climate matters beyond conventional goals (Nagaj et al., 2024).

The strong positive effect of water stress, particulate pollution, Heating Degree Days, and nitrous oxide emissions underscores that emissions from buildings are a phenomenon within larger ecological and climatic stress dynamics. Countries with high levels of water stress and pollution, particularly Southern and Eastern Europe, should focus on adapting and making energy more efficient in a manner that tackles climate change. The findings above also suggest that traditional energy efficiency measures might be insufficient when fundamental factors drive energy demand up. Therefore, there is a broader policy challenge that encompasses not just better technologies within buildings but also environmental factors that accentuate emissions (Sechi et al., 2022).

Machine learning outcomes again confirm that "the electricity mix plays a critical role in which GHG emissions." The importance of transitioning away from coal and promoting renewable energy sources in the energy mix becomes clearer as a result, as "coal-based generation" ranks as a highly influential determinant. Despite progress made by the EU in curtailing its usage, regions with highly polluting nations are still identifiable as being driven by fossil fuel-dominated energy sources. Improving enforcement of the Renewable Energy Directive and developing more interconnections with surrounding regions can effectively address emissions within the sector. Evidence also shows that energy dependence and energy intensity contribute significantly to emissions, proving once again the relevance of regional cooperation on energy security and efficiency issues (Cazcarro et al., 2022).

The clustering outcome shows a clear advantage for differentiated policy needs based on regional structural characteristics. Emissions are driven jointly for high-emission clusters, involving agriculture, air pollution, fossil fuel consumption, and inefficiencies within infrastructures (Sechi et al., 2022). Member States within these groups would be best served through comprehensive and simultaneous efforts related to electricity sources, rates of building renovation, and pollution abatement. Members within low-emission clusters with factors reflecting larger use of renewables, large forest areas, and more advantageous climatic conditions would be served better by policies addressing and preserving these superior conditions and helping make them more resilient against variable climatic changes (Fotiou et al., 2024; Moran et al., 2022).

The classification of regions into groups with large numbers of cooling and heating degree days also implies that there should be a more prominent role for climate-adaptive architectural designs within the Renovation Wave and changes to EPBD, still not sufficiently addressed within modern European policy frameworks.

A central methodological consideration that arises here regards the low level of emissions inertia shown by the Dynamic Panel approach and implies that emissions could be significantly reduced within a short period following policy actions. Nevertheless, also significant here is that GMM estimation shows an issue within instrument validity, pointing out challenges for identifying causal links within a global and interdependent environment and instead indicating that empirical research should be given a central role within policy-making and should not primarily focus on interpreting causality implications (Cazcarro et al., 2022).

In summary, the findings from these three methods encourage an overall integrated strategy of decarbonisation based on the structural heterogeneity of the European Union. Climate change policies within the sector of buildings have to address, at the same time, energy sector change, adaptation to climate change, environmental conservation, and structural change within various economies. The cumulative evidence from econometric, machine learning, and clustering analysis emphasises that it will never be possible by single measures alone to attain the necessary level of emission cuts. It will be necessary instead to apply multi-sectoral and country- and structure-specifically composed climate policy strategies for reaching the ambitious long-term aims of climate neutrality adopted within the EU framework (Fotiou et al. 2024; Giannelos et al. 2023; Nagaj et al. 2024).

9. Limitations and Directions for Future Research

Although the scope and scientific nature of this research are wide and rigorous, some limitations exist that must be noted in order to explain the research approach and point towards future research directions. These limitations do not affect the validity of the findings but instead form a basis for interpretation.

First, the study is based on secondary data that is more macro in orientation, sourced from World Bank-type datasets. As is common in cross-national research that is longitudinal in focus, there are gaps in data

coverage for Member States in Europe for different years. This means that there are missing data for specific variables for specific years in the dataset. These missing data were handled through the use of the unbalanced panel method that allows for the utilisation of data that is available without interpolation or imputation, in keeping with current debates in panel data research that have argued for trade-offs between data completeness and estimation bias (Gao et al., 2025).

Although this method is useful in that it does not rely on data that is created through interpolation or assumptions in imputation, there may be limitations in its utilisation in research that is focused on country-specific dynamics for shorter time series, in that there may be interruptions in time series that make country data incomplete (Mele et al., 2025).

Thirdly, variables are mostly applied in their raw form without any log or nonlinear transformation. The objective of this choice of methodology is to ensure interpretability and consistency of results from all the empirical elements of this research project, such as panel econometric models, machine learning models, and clustering analysis. In fact, this is very common in modern mixed-method studies. Although this improves cross-methodological comparison of results, this choice of methodology makes it difficult to interpret results related to elasticities and any possible effects of scale. Future studies may consider other transformation methods, for instance, log transformation (Hoa et al., 2024).

Fourth, although a variety of panel estimation methods were used, including Random Effects, Fixed Effects, Dynamic Panel GMM, and Weighted Least Squares, concerns of potential endogeneity cannot be fully eliminated. Exogenous instruments from outside were not considered because of difficulties related to finding appropriate instruments that are truly exogenous within the very complex EU context of the environment and climate (Gul et al., 2025). Dynamic Panel GMM was mainly used as a robustness check with internal instruments. However, issues related to instrument accumulation and over-identifications are noted to be applicable to practical applications of GMM (Gul et al., 2025). Consequently, results from GMM should be treated with care and with more focus on results from FE and WLS.

Fifth, the findings are based on aggregate country-level data, which could dampen the existence of considerable subnational variation. Climate patterns at the regional level, urbanisation rates, building stocks, and the implementation of policies can strongly impact emissions in the buildings sector and can be heterogeneous at the country level (Mele et al., 2025). It will be valuable for future research to explore country-level findings at a subnational level, such as a region or a city level, to shed light on the underlying drivers of emissions.

Sixth, although the proposed ESG-based analysis structure provides an effective way of conceptualising factors related to the environment, social, and governance domain, governance factors are addressed only indirectly through proxy variables that relate to the energy system and emission performance. The lack of direct governance and policy performance data in this area has not been addressed in recent empirical studies that focus on ESG (Drago et al., 2025; Caprioli et al., 2024), and this should be addressed in future studies as data becomes more readily available (Montero et al., 2025).

Lastly, although the combination of machine learning methods and clustering analysis improves their robustness and explainability, these methods are mostly exploratory in nature, supplementary to the econometric analysis. In fact, machine learning methods are mostly developed for predictive purposes, for uncovering non-linear patterns, rather than for causal analysis, while clustering analysis detects structural similarities without suggesting causality (Cerqua et al., 2025; Guo et al., 2024). As such, findings from these analyses can be considered supplementary to findings from econometric analysis. Future studies might investigate causal machine learning or dynamic clustering analysis to uncover transition patterns in regimes and emissions.

9.1. Descriptive Nature of the Empirical Analysis

The empirical strategy in this research does not focus on investigating possible causality between the set of explanatory variables and the variable of carbon dioxide emissions in the building sector. The model framework adopted in this research is intended to investigate statistical associations, structural patterns, and cross-country similarities by combining panel econometrics, machine learning algorithms, and clustering methods. Because it is not based on an empirical strategy that focuses on instrumentally identifying causality, it is not appropriate to interpret these empirical outcomes in terms of causality.

Panel data estimators are employed to analyse the average association and co-variation of emissions with environment, structure, and climate variables in the member states of the European Union, accounting for country-specific heterogeneity and time dynamics. Machine learning algorithms play a secondary and exploratory analytical role in determining the importance of correlated drivers of emissions, whose association could not be captured in an econometric model. Clustering algorithms are also utilised to categorise member states with similar attributes, providing a typological interpretation of emissions, as opposed to causal categorisation.

All the empirical results are interpreted from a descriptive and storytelling point of view. The analysis yields evidence on the co-variation of the variables across countries and over time, focusing on the patterns and structural arrangements that are associated with the various emission patterns. The policy implications are treated with the necessary care and focus on the correlative drivers, structural risks, and country profiles, rather than direct policy effects and effectiveness. In this way, the interpretation of the empirical findings is guaranteed to stay within the bounds of the data and methods used, while still providing useful insights into the emission dynamics in the European building sector.

Although causal inference is not within the scope of this research, it is worth noting that while causal inference is not possible here, the results of the various analyses performed—both econometric and machine learning, as well as clustering—provide enough signals that can be used for identifying risk factors and country profiles that can be prioritised.

10. Conclusions

The research describes a comprehensive and multi-criteria analysis on factors influencing carbon dioxide emissions from the building sector (CBE) in all 27 Member States within the European Union from 2005 to 2023. It uses a combination of econometric modelling, learning algorithms, and clustering methods. From a methodological perspective, this research shows that it is necessary to combine multiple tools because these tools alone cannot elaborate on and explain CBE factors.

The result shows that a set of variables affects CBE. The value added within Agriculture, Forestry, and Fishing activities (AFFV) and forest area (FRST) have highly Negative associations with CBE. It implies that rural structure and large natural carbon sinks correspond with low emission rates. It suggests that emission rates are moderated by demographic and land distribution variables.

Variables on water stress (WSTR), PM2.5, N2OP, and HDD show highly Significant and Positive associations with CBE. It suggests that emission rates are determined by and associated with environmental stress, Climatic factors, and pollution/energy interactions. Diagnostics result shows that the so-called Fixed Effects and Weighted Least Squares methods have stronger estimation capabilities and have addressed modelling deficiencies on heterogeneity and heteroscedasticity across Member States.

Although persistence and endogeneity were considered with DGMM, it failed as an ordering solution due to an invalid instrument set. It suggests that analysis based on CBE and comparable research would have methodological limitations in ascertaining causality and fact due to global interdependency and spatial implications.

Machine learning expands and strengthens related research on CBE based on its Non-Linear associations. It ranks in importance among predicted variables. The K-Nearest Neighbours algorithm and resultant feature alignment show more consistency and closeness among observation sets. It identifies the main CBE drivers as generation of electricity via coal consumption, Water stress, Air pollution, and Temperature. It shows that variables influencing CBE would be influenced and moderated by structural elements, energy interdependences with Climatic intensification.

The clustering analysis brings out a marked differentiation among Member States within the EU. Countries belonging to high-emission clusters share common characteristics associated with agriculture, pollution, and fossil fuel usage, and sometimes energy inefficiency. By contrast, low-emission clusters have common advantageous characteristics, including large forested areas or more congenial climatic conditions that make energy conservation easier.

The implications and implications trickle down to emphasising that there should be specific country plans within an overall strategy towards making Europe a zero-carbon economy. At an overall level, it lends an integrated and statistically sound platform for understanding factors resisting the reduction of CO_2 emissions

within Europe's building sector. By pointing out that emissions occur owing to a joint effect of economic structure, climate factors, and energy characteristics, Europe should be made aware that there should be comprehensive policy approaches instead of uniform thoughts on becoming a zero-carbon economy.

Acknowledgements Not Applicable

Author Contributions Conceptualisation, M.M., A.C., F.A., A.L.; methodology, M.M., A.C., F.A., A.L.; validation, M.M., A.C., F.A., A.L.; formal analysis, M.M., A.C., F.A., A.L.; investigation, M.M., A.C., F.A., A.L.; resources, M.M., A.C., F.A., A.L.; data curation, M.M., A.C., F.A., A.L.; writing—original draft preparation, M.M., A.C., F.A., A.L.; writing—review and editing, M.M., A.C., F.A., A.L.; supervision, M.M., A.C., F.A., A.L.; project administration, M.M., A.C., F.A., A.L.. All authors have read and agreed to the published version of the manuscript.

Funding Not Applicable

Data Availability The data is available at the following link: https://data360.worldbank.org/en/search

Declarations

Competing Interests The authors declare no competing interests.

References

Abualhaj, M. M., Abu-Shareha, A. A., Shambour, Q. Y., Alsaaidah, A., Al-Khatib, S. N., & Anbar, M. (2024). Customized K-nearest neighbors' algorithm for malware detection. International Journal of Data & Network Science, 8(1).

Akakpo, S., Dambra, P., Paz, R., Smyth, T., Torre, F., & Yu, C. (2024). Optimization of the K-Nearest Neighbor Algorithm to Predict Bank Churn. Statistics, Optimization & Information Computing, 12(5), 1397-1408.

Almusaed, A., Yitmen, I., Myhren, J. A., & Almssad, A. (2024). Assessing the impact of recycled building materials on environmental sustainability and energy efficiency: a comprehensive framework for reducing greenhouse gas emissions. Buildings, 14(6), 1566.

Aquino, A., Bassetti, M., Grasso, E., & Martini, F. (2024, November). Energy Efficiency of the Office Buildings in Italy: Insights for the European Taxonomy. In Journal of Physics: Conference Series (Vol. 2893, No. 1, p. 012043). IOP Publishing.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., & nigo Perona, I. (2014). An extensive comparative study of cluster validity indices. Pattern Recognition.

Asdrubali, F., Berardi, U., & Stasi, R. (2025). The impact of the building sector. In Sustainability Certifications, Labels and Tools in the Built Environment (pp. 4-21). Routledge.

Aste, N., Del Pero, C., & Leonforte, F. (2022). Toward building sector energy transition. In Handbook of Energy Transitions (pp. 127-150). CRC Press.

Attia, S., Santos, M. C., Al-Obaidy, M., & Baskar, M. (2021, October). Leadership of EU member States in building carbon footprint regulations and their role in promoting circular building design. In IOP Conference Series: Earth and Environmental Science (Vol. 855, No. 1, p. 012023). IOP Publishing.

Awad, A., & Warsame, M. H. (2022). The poverty-environment nexus in developing countries: Evidence from heterogeneous panel causality methods, robust to cross-sectional dependence. Journal of Cleaner Production, 331, 129839.

Babbar, H., Rani, S., Sah, D. K., AlQahtani, S. A., & Kashif Bashir, A. (2023). Detection of android malware in the Internet of Things through the K-nearest neighbor algorithm. Sensors, 23(16), 7256.

Balaras, C. A., Dascalaki, E. G., Patsioti, M., Droutsa, K. G., Kontoyiannidis, S., & Cholewa, T. (2023). Carbon and greenhouse gas emissions from electricity consumption in European Union Buildings. Buildings, 14(1), 71.

Bărbulescu, A. (2024). Statistical Analysis and Modeling of the CO_2 Series Emitted by Thirty European Countries. Climate, 12(3), 34.

Bardulis, A., Purvina, D., Bardule, A., & Lazdins, A. (2023). Potential role of tree introduction in agricultural land to reduce greenhouse gas emissions. In 22nd International Scientific Conference" Engineering for Rural Development": proceedings:[Jelgava, Latvia,] May 24-26, 2023 (Vol. 22, pp. 196-203).

Bezić, H., Mance, D., & Balaž, D. (2022). Panel evidence from EU countries on CO_2 emission indicators during the Fourth Industrial Revolution. Sustainability, 14(19), 12554.

Biancalani, F., Gnecco, G., Metulini, R., & Riccaboni, M. (2024). The impact of the European Union emissions trading system on carbon dioxide emissions: a matrix completion analysis. Scientific reports, 14(1), 19676.

Boca, M. C., Bungau, C. C., & Hanga-Farcas, I. F. (2025). Climate-Conscious Sustainable Practices in the Romanian Building Sector. Buildings, 15(12), 2106.

Bonar, D. J. (2024). Do strict environmental policies in European countries reduce CO_2 emissions?. Przegląd Statystyczny. Statistical Review, 71(1), 1-22.

Braungardt, S., Bei der Wieden, M., & Kranzl, L. (2025). EU emissions trading in the buildings sector–an ex-ante assessment. Climate Policy, 25(2), 208-222.

Cámara-Aceituno, J., Hermoso-Orzáez, M. J., Terrados-Cepeda, J., Rivadeneira-Zambrano, A., Mena-Nieto, Á., Golpe, A. A., & Garcia-Ramos, J. E. (2025). Exploring the driving forces of CO_2 emissions in the European Union. Open Research Europe, 5, 132.

Caprioli, S., Foschi, J., Crupi, R., & Sabatino, A. (2024). Denoising ESG: quantifying data uncertainty from missing data with Machine Learning and prediction intervals. arXiv preprint arXiv:2407.20047.

Cazcarro, I., García-Gusano, D., Iribarren, D., Linares, P., Romero, J. C., Arocena, P., ... & Cadarso, M. Á. (2022). Energy-socio-economic-environmental modelling for the EU energy and post-COVID-19 transitions. Science of the Total Environment, 805, 150329.

Cerqua, A., Letta, M., & Pinto, G. (2025). On the (mis) use of machine learning with panel data. Oxford Bulletin of Economics and Statistics.

Charlier, D., Fodha, M., & Kirat, D. (2023). Residential CO_2 emissions in Europe and carbon taxation: a country-level assessment. The Energy Journal, 44(5), 187-206.

Chen, L., Zhang, Y., Chen, Z., Dong, Y., Jiang, Y., Hua, J., ... & Yap, P. S. (2024). Biomaterials technology and policies in the building sector: a review. Environmental Chemistry Letters, 22(2), 715-750.

Choi, Y., An, N., Hong, S., Cho, H., Lim, J., Han, I. S., ... & Kim, J. (2022). Time-series clustering approach for training data selection of a data-driven predictive model: Application to an industrial bio 2, 3-butanediol distillation process. Computers & Chemical Engineering, 161, 107758.

Chovancová, J., Petruška, I., & Litavcová, E. (2021). Dependence of CO_2 emissions on energy consumption and economic growth in the European Union: A panel threshold model. Economics and Environment, 78(3), 73-89.

Cialani, C., & Mortazavi, R. (2021). Sectoral analysis of club convergence in EU countries' CO_2 emissions. Energy, 235, 121332.

Costantiello, A., Laureti, L., Quarto, A., & Leogrande, A. (2025). Methane Emissions in the ESG Framework at the World Level. Methane, 4(1), 3.

Cubuk, G. (2023). Comparison of emissions gap reports by building sector between 2014–2022 through data analysis. Journal of Environmental Protection and Ecology, 24(2), 397-407.

Dai, Q., Liu, X. Y., Sun, F. Y., & Ren, F. R. (2024). Ensemble intelligence prediction algorithms and land use scenarios to measure carbon emissions of the Yangtze River Delta: A machine learning model based on Long Short-Term Memory. PloS One, 19(12), e0311441.

Dinca, G., Barbuța, M., Negri, C., Dinca, D., & Model, L. S. (2022). The impact of governance quality and educational level on environmental performance. Front. Environ. Sci, 10, 1-15.

Dominguez, C., Kakkos, E., Gross, D., Hischier, R., & Orehounig, K. (2024). Renovated or replaced? Finding the optimal solution for an existing building considering cumulative CO_2 emissions, energy consumption and costs–A case study. Energy and Buildings, 303, 113767.

Doran, N. M., Bădîrcea, R. M., Jianu, E., Antoniu, M. E., Ciobanu, R. M., & Ciobanu, Ș. C. F. (2025). Unveiling CO_2 Emission Dynamics Under Innovation Drivers in the European Union. Sustainability, 17(8), 3463.

Doryń, W., & Wawrzyniak, D. (2024). Heterogeneity in air pollution levels and their techno-economic determinants: a cluster analysis of the EU–27.

Drago, C., Arnone, M., & Leogrande, A. (2025). A Machine Learning and Panel Data Analysis of N2O Emissions in an ESG Framework. Sustainability, 17(10), 4433.

Drago, C., Costantiello, A., Arnone, M., Anobile, F., & Leogrande, A. (2025). Decoding Energy Consumption in the ESG Era: Panel Data Evidence and Machine Learning Insights.

Dragonetti, L., Papadaki, D., Mazzoli, C., Monacelli, A., Assimakopoulos, M. N., & Ferrante, A. (2025). Circular deep renovation versus demolition with reconstruction: Environmental and financial evaluation to support decision making in the construction sector. Energy and Buildings, 336, 115610.

Elizabeth Yancey, R., Xin, B., & Matloff, N. (2021). Modernizing k-nearest neighbors. Stat, 10(1), e335.

Erdogan, S. (2021). Dynamic nexus between technological innovation and building sector carbon emissions in the BRICS countries. Journal of Environmental Management, 293, 112780.

Famiglietti, J., Madioum, H., & Motta, M. (2023). Developing a New Data-Driven LCA Tool at the Urban Scale: The Case of the Embodied Environmental Profile of the Building Sector. Sustainability, 15(15), 11518.

Fotiou, T., Fragkos, P., & Zisarou, E. (2024). Decarbonising the EU Buildings| Model-Based Insights from European Countries. Climate, 12(6), 85.

Gan, L., Liu, Y., & Cai, W. (2023). Carbon neutral projections of public buildings in China under the shared socioeconomic pathways: A tertiary industry perspective. Environmental Impact Assessment Review, 103, 107246.

Gao, J., Liu, F., Peng, B., & Yan, Y. (2025). Panel Data Estimation and Inference: Homogeneity versus Heterogeneity. arXiv preprint arXiv:2502.03019.

Gholipour, H. F., Arjomandi, A., & Yam, S. (2022). Green property finance and CO_2 emissions in the building industry. Global Finance Journal, 51, 100696.

Giannelos, S., Bellizio, F., Strbac, G., & Zhang, T. (2024). Machine learning approaches for predictions of CO_2 emissions in the building sector. Electric Power Systems Research, 235, 110735.

Giannelos, S., Moreira, A., Papadaskalopoulos, D., Borozan, S., Pudjianto, D., Konstantelos, I., ... & Strbac, G. (2023). A machine learning approach for generating and evaluating forecasts on the environmental impact of the buildings sector. Energies, 16(6), 2915.

Gul, R., Hussain, S., Shaikh, A., & Naifar, N. (2025). Examining the impact of green finance and green innovation on sustainable economic development using machine learning methods. Discover Sustainability, 6(1), 1333.

Guo, Y., Guan, C., & Ma, J. (2024). Exioml: Eco-economic dataset for machine learning in global sectoral sustainability. arXiv preprint arXiv:2406.09046.

Guo, Z., & Luo, Y. (2024). A Building Carbon Emission Prediction Model Based on Optimized Machine Learning Model. Innovative Applications of AI, 1(4), 70-83.

Hasler, C., & Tillé, Y. (2016). Balanced k-nearest neighbour imputation. Statistics, 50(6), 1310-1331.

Hassan, O. A., & Rezaei, H. (2025). The Load-Bearing System from the Perspective of Sustainable Building. Green and Low-Carbon Economy, 3(3).

Heinz, E., Sovacool, B. K., Kwan, T., & Petit, V. (2025). The Barriers and Drivers to Decarbonization of the Building Sector in the United States: Qualitative Insights from Boston and Phoenix. Building and Environment, 114032.

Hemmati, M., Messadi, T., Gu, H., Seddelmeyer, J., & Hemmati, M. (2024). Comparison of embodied carbon footprint of a mass timber building structure with a steel equivalent. Buildings, 14(5), 1276.

Hoa, P. X., Xuan, V. N., & Thu, N. T. P. (2024). Factors affecting carbon dioxide emissions for sustainable development goals–New insights into six asian developed countries. Heliyon, 10(21).

Hsu, A., Wang, X., Tan, J., Toh, W., & Goyal, N. (2022). Predicting European cities' climate mitigation performance using machine learning. Nature Communications, 13(1), 7487.

Hu, S., Jiang, Y., Yang, X., Pan, Y., Rong, X., Hao, B., ... & Yan, D. (2025). Ecological Pathway to Achieve Carbon Neutrality in China's Building Sector. Engineering.

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. Heliyon, 11(2).

Iqbal, A., Zhang, W., & Jahangir, S. (2025). Building a Sustainable Future: The Nexus Between Artificial Intelligence, Renewable Energy, Green Human Capital, Geopolitical Risk, and Carbon Emissions Through the Moderating Role of Institutional Quality. Sustainability, 17(3), 990.

Iwanicz-Drozdowska, M., Lament, M., & Witkowski, B. (2025). ESG Performance and Economic Growth in Europe. Sustainable Development.

Junda, E., & Málaga-Chuquitaype, C. (2025). Life cycle impacts of structural deterioration and seismic events on cross-laminated timber buildings. Journal of Building Engineering, 104, 112282.

Kadrić, D., Aganovic, A., Martinović, S., Delalić, N., & Delalić-Gurda, B. (2022). Cost-related analysis of implementing energy-efficient retrofit measures in the residential building sector of a middle-income country–A case study of Bosnia and Herzegovina. Energy and Buildings, 257, 111765.

Kartal, M. T., Kirikkaleli, D., & Pata, U. K. (2024). Role of environmental policy stringency on sectoral CO_2 emissions in EU-5 countries: disaggregated level evidence by novel quantile-based approaches. Energy & Environment, 0958305X241241026.

Kayakuş, M., Terzioğlu, M., Erdoğan, D., Zetter, S. A., Kabas, O., & Moiceanu, G. (2023). European Union 2030 carbon emission target: The case of Turkey. Sustainability, 15(17), 13025.

Koengkan, M., Fuinhas, J. A., Teixeira, M., Kazemzadeh, E., Auza, A., Dehdar, F., & Osmani, F. (2022). The capacity of battery-electric and plug-in hybrid electric vehicles to mitigate CO_2 emissions: macroeconomic evidence from european union countries. World Electric Vehicle Journal, 13(4), 58.

Kosowski, P. (2024). From Fossil Fuels to Renewables: Clustering European Primary Energy Production from 1990 to 2022. Energies, 17(22), 5596.

Lenssen, L., & Schubert, E. (2024). Medoid Silhouette clustering with automatic cluster number selection. Information Systems, 120, 102290.

Levada, A. L. M., Nielsen, F., & Haddad, M. F. C. (2024). Adaptive $ k $-nearest neighbor classifier based on the local estimation of the shape operator. arXiv preprint arXiv:2409.05084.

Li, S., Siu, Y. W., & Zhao, G. (2021). Driving factors of CO_2 emissions: further study based on machine learning. Frontiers in Environmental Science, 9, 721517.

Li, X., Li, Y., Zhou, H., Fu, Z., Cheng, X., & Zhang, W. (2023). Research on the carbon emission baselines for different types of public buildings in a northern cold areas city of China. Buildings, 13(5), 1108.

Li, Y., Chen, H., Yu, P., & Yang, L. (2024). The application and evaluation of the LMDI method in building carbon emissions analysis: A comprehensive review. Buildings, 14(9), 2820.

Liu, Y., Wu, Y., & Zhu, X. (2024). Industrial clusters and carbon emission reduction: evidence from China. The Annals of Regional Science, 73(2), 557-597.

Magaletti, N., Notarnicola, V., Di Molfetta, M., & Leogrande, A. (2025). Decarbonizing the building sector: the integrated role of environmental, social, and governance indicators. Buildings, 15(19), 3601.

Mailagaha Kumbure, M., & Luukka, P. (2022). A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. Granular Computing, 7(3), 657-671.

Mandel, T., Kranzl, L., Popovski, E., Sensfuß, F., Müller, A., & Eichhammer, W. (2023). Investigating pathways to a net-zero emissions building sector in the European Union: what role for the energy efficiency first principle?. Energy Efficiency, 16(4), 22.

Martinho, M., Fernandes, J., Gomes, R., Lourador, P., & Ferrão, P. (2025). A comprehensive tool for embodied carbon quantification and supporting mitigation strategies in the Building sector. Discover Applied Sciences, 7(7), 781.

Meena, C. S., Kumar, A., Jain, S., Rehman, A. U., Mishra, S., Sharma, N. K., ... & Eldin, E. T. (2022). Innovation in green building sector for sustainable future. Energies, 15(18), 6631.

Mele, M., Costantiello, A., Anobile, F., & Leogrande, A. (2025). Determinants of Building-Sector CO_2 Emissions in the EU: A Combined Econometric and Machine Learning Approach.

Mohammed, K. S., Pata, U. K., Serret, V., & Kartal, M. T. (2024). The role of renewable energy and carbon dioxide emissions on the ESG market in European Union. Managerial and Decision Economics, 45(7), 5146-5158.

Mohammed, M. A., Abd Ghani, M. K., Hamed, R. I., Mostafa, S. A., Ibrahim, D. A., Jameel, H. K., & Alallah, A. H. (2017). Solving vehicle routing problem by using improved K-nearest neighbor algorithm for best solution. Journal of Computational Science, 21, 232-240.

Montero, J. M., Valls Martínez, M. D. C., & Santos-Jaén, J. M. (2025). The impact of board gender diversity on green building practices: Moving beyond traditional linear and logistic specifications. Corporate Social Responsibility and Environmental Management, 32(2), 1779-1830.

Moran, D., Pichler, P. P., Zheng, H., Muri, H., Klenner, J., Kramel, D., ... & Gurney, K. R. (2022). Estimating CO_2 emissions for 108 000 European cities, Earth Syst. Sci. Data, 14, 845–864.

Morelli, C., Maranzano, P., & Otto, P. (2025). Spatiotemporal clustering of GHGs emissions in Europe: exploring the role of spatial component. arXiv preprint arXiv:2503.11909.

Mushtaq, M., Ahmed, S., Abbas, A., & Fahlevi, M. (2024). Impact of urbanization on environmental eminence: Moderating role of renewable energy. International Journal of Energy Economics and Policy, 14(2), 244-257.

Myint, N. N., Shafique, M., Zhou, X., & Zheng, Z. (2025). Net zero carbon buildings: A review on recent advances, knowledge gaps and research directions. Case Studies in Construction Materials, e04200.

Nader, Y., Sixt, L., & Landgraf, T. (2022, June). DNNR: Differential nearest neighbors regression. In International Conference on Machine Learning (pp. 16296-16317). PMLR.

Nagaj, R., Gajdzik, B., Wolniak, R., & Grebski, W. W. (2024). The impact of deep decarbonization policy on the level of greenhouse gas emissions in the European Union. Energies, 17(5), 1245.

Nichifor, B., Zait, L., & Turcu, O. (2025). Renewable Investments, Environmental Spending, and Emissions in Eastern Europe: A Spatial-Economic Analysis of Management and Policy Decisions Efficiency. Sustainability, 17(7), 3010.

Nigmatullaeva, G., Ibragimova, F., Dekhkanova, N., Umarov, A., & Sadikov, A. (2025). Renewable energy, private sector development, and CO_2 emissions: evidence from early demographic dividend countries. International Journal of Energy Economics and Policy, 15(5), 705-713.

Nurgaliyeva, K. (2025). Evaluating the Impact of ESG on Regional Development in Kazakhstan: Empirical Analysis. Eurasian Journal of Economic and Business Studies, 69(1), 5-17.

Okunevičiūtė Neverauskienė, L., Dirma, V., Tvaronavičienė, M., & Danilevičienė, I. (2025). Assessing the Role of Renewable Energy in the Sustainable Economic Growth of the European Union. Energies, 18(4), 760.

ÖZEN, E., HAZAR, A., GRIMA, S., MISTREAN, L., & SAÇKES, E. (2025). IX. INTERNATIONAL APPLIED SOCIAL SCIENCES CONGRESS (C-IASOS 2025).

Ozturk Kiyak, E., Ghasemkhani, B., & Birant, D. (2023). High-Level K-Nearest Neighbors (HLKNN): A supervised machine learning model for classification analysis. Electronics, 12(18), 3828.

Papachatzis, K. (2024). Machine learning-based price prediction for thermal insulation materials: a holistic approach integrating thermophysical, technical, and environmental attributes in the Greek construction market. Energy and Buildings, 324, 114899.

Papadas, D., Ghosh, B., & Kostakis, I. (2024). Investigating the role of energy mix and sectoral decomposition on environmental sustainability in selected European countries. Development and Sustainability in Economics and Finance, 1, 100001.

Papangelopoulou, M. D., Alexakis, K., & Askounis, D. (2025). Assessment Methods for Building Energy Retrofits with Emphasis on Financial Evaluation: A Systematic Literature Review. Buildings, 15(14), 2562.

Park, H. G., Shin, K. S., & Kim, J. C. (2025). Efficient Clustering Method for Graph Images Using Two-Stage Clustering Technique. Electronics, 14(6), 1232.

Pesaran, M. H. (2021). General diagnostic tests for cross-sectional dependence in panels. Empirical economics, 60(1), 13-50.

Petrescu, A. M. R., Peylin, P., Matthews, B., Dentener, F., Balkovic, J., Bastrikov, V., ... & Walther, S. (2023). The consolidated European synthesis of CO_2 emissions and removals for the European Union and United Kingdom: 1990–2020. Earth System Science Data, 15(10), 4295-4370.

Petruška, I., Litavcová, E., & Chovancová, J. (2022). Impact of renewable energy sources and nuclear energy on CO_2 emissions reductions—the case of the EU countries. Energies, 15(24), 9563.

Piccardo, C., Alam, A., & Hughes, M. (2021). The Potential Contribution of Wood in Green Building Certifications: prospects in sustainable residential buildings. Architectural Research in Finland, 5(1), 130-146.

Rakhshan, K., Daneshkhah, A., & Morel, J. C. (2023). Stakeholders' impact on the reuse potential of structural elements at the end-of-life of a building: A machine learning approach. Journal of Building Engineering, 70, 106351.

Raycheva, I. Economic Growth, Energy Consumption and CO_2 Emissions in the European Union. A Panel Data Analysis. In LIMEN 2023/9–Leadership, Innovation, Management and Economics: Integrated Politics of Research-CONFERENCE PROCEEDINGS (pp. 245-251). Udruženje ekonomista i menadžera Balkana.

Reyes, N., Rodríguez, B., Wiegand, E., Zilic, F., Ramage, M., Bukauskas, A., ... & Gin, Y. (2021). Achieving zero carbon emissions in the construction sector: The role of timber in decarbonising building structures.

Rheude, F., & Röder, H. (2022). Estimating the use of materials and their GHG emissions in the German building sector. Cleaner Environmental Systems, 7, 100095.

Rietig, K. (2021). Multilevel reinforcing dynamics: Global climate governance and European renewable energy policy. Public Administration, 99(1), 55-71.

Roca Reina, J. C., Carlsson, J., Volt, J., & Toleikyte, A. (2025). Alternatives for Decarbonising High-Temperature Heating Facilities in Residential Buildings. Energies, 18(2), 235.

Rossbroich, J., Durieux, J., & Wilderjans, T. F. (2022). Model selection strategies for determining the optimal number of overlapping clusters in additive overlapping partitional clustering. Journal of Classification, 39(2), 264-301.

Sadhukhan, S., & Yadav, V. K. (2023). Forecasting, capturing and activation of carbon-dioxide (CO $ _2 $): Integration of Time Series Analysis, Machine Learning, and Material Design. arXiv preprint arXiv:2307.14374.

Sarıca, K., Harputlugil, G. U., İnaner, G., & Kollugil, E. T. (2023). Building sector emission reduction assessment from a developing European economy: A bottom-up modelling approach. Energy Policy, 174, 113429.

Sasmita, N. R., Khairul, M., Sofyan, H., Kruba, R., Mardalena, S., Dahlawy, A., ... & Maula, A. W. (2023). A Statistical Clustering Approach: Mapping Population Indicators Through Probabilistic Analysis in Aceh Province, Indonesia. Infolitika Journal of Data Science, 1(2), 63-72.

Sechi, S., Giarola, S., & Leone, P. (2022). Taxonomy for Industrial Cluster Decarbonization: An Analysis for the Italian Hard-to-Abate Industry. Energies, 15(22), 8586.

Seyedabadi, M. R., Karrabi, M., Shariati, M., Karimi, S., Maghrebi, M., & Eicker, U. (2024). Global building life cycle assessment: Comparative study of steel and concrete frames across European Union, USA, Canada, and Australia building codes. Energy and Buildings, 304, 113875.

Sharmina, M., Pappas, D., Scott, K., & Gallego-Schmid, A. (2023). Impact of circular economy measures in the European Union built environment on a net-zero target. Circular Economy and Sustainability, 3(4), 1989-2008.

Song, X., Zhai, S., & Zhou, N. (2024). The carbon emissions from public buildings in China: a systematic analysis of evolution and spillover effects. Sustainability, 16(15), 6622.

Suproń, B., & Myszczyszyn, J. (2024). Exploring the Dynamic Relationships between Agricultural Production and Environmental Pollution: Evidence from a GMM-SYS Model in the Three Seas Initiative (3SI). Sustainability, 16(9), 3748.

Tudor, C., Sova, R., Stamatiou, P., Vlachos, V., & Polychronidou, P. (2025). Future-Proofing EU-27 Energy Policies with AI: Analyzing and Forecasting Fossil Fuel Trends. Electronics, 14(3), 631.

Tzeiranaki, S. T., Bertoldi, P., Economidou, M., Clementi, E. L., & Gonzalez-Torres, M. (2023). Determinants of energy consumption in the tertiary sector: Evidence at European level. Energy Reports, 9, 5125-5143.

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1), 6256.

Usman, M., & Jahanger, A. (2021). Heterogeneous effects of remittances and institutional quality in reducing environmental deficit in the presence of EKC hypothesis: a global study with the application of panel quantile regression. Environmental Science and Pollution Research, 28(28), 37292-37310.

Vagnini, C., Canal Vieira, L., Longo, M., & Mura, M. (2025). Regional drivers of industrial decarbonization: a spatial econometric analysis of 238 EU regions between 2008 and 2020. Regional Studies, 59(1), 2380369.

Velemir Radović, M., Nikolic, D., & Jovanović, S. (2024). ENERGY EFFICIENCY IN THE BUILDING SECTOR IN SERBIA-AN OVERVIEW.

Velentzas, P., Moutafis, P., & Mavrommatis, G. (2020, November). An improved GPU-based algorithmfor processing the k nearest neighbor query. In Proceedings of the 24th Pan-Hellenic Conference on Informatics (pp. 372-375).

Vigna, I., Lollini, R., & Pernetti, R. (2021). Assessing the energy flexibility of building clusters under different forcing factors. Journal of Building Engineering, 44, 102888.

Wei, J., Shi, W., Ran, J., Pu, J., Li, J., & Wang, K. (2023). Exploring the driving factors and their spatial effects on carbon emissions in the building sector. Energies, 16(7), 3094.

Xia, Y., Yang, Z., Jiang, X., & Wang, H. (2024). The road to carbon neutrality in China's building sector. Iscience, 27(9).

Xin, L., Li, S., Rene, E. R., Lun, X., Zhang, P., & Ma, W. (2023). Prediction of carbon emissions peak and carbon neutrality based on life cycle CO_2 emissions in megacity building sector: Dynamic scenario simulations of Beijing. Environmental Research, 238, 117160.

Yakymchuk, A., & Rataj, M. A. (2025). Economic Analysis of Fossil CO_2 Emissions: A European Perspective on Sustainable Development. Energies, 18(8), 2106.

Yang, S., Yang, D., Shi, W., Deng, C., Chen, C., & Feng, S. (2023). Global evaluation of carbon neutrality and peak carbon dioxide emissions: Current challenges and future outlook. Environmental Science and Pollution Research, 30(34), 81725-81744.

Yao, Z., Li, W., & Pang, Y. (2024). Environmental modeling of impacts of agricultural land changes using Markov chain and machine learning (case study: Shanghai metropolis, China). International Agrophysics, 38(4), 353-371.

You, K., Li, Y., Cai, W., Zhang, L., Liu, Z., Feng, W., & Wei, Y. M. (2025). Mitigating emissions and costs through demand-side solutions in Chinese residential buildings. Nature communications, 16(1), 7358.

Zangana, H. M., & Abdulazeez, A. M. (2023). Developed clustering algorithms for engineering applications: A review. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 4(2), 160-182.

Żelazna, A., & Pawłowski, A. (2025). Review of the Role of Heat Pumps in Decarbonization of the Building Sector. Energies, 18(13), 3255.

Zhang, T., & Lin, G. (2021). Generalized k-means in GLMs with applications to the outbreak of COVID-19 in the United States. Computational Statistics & Data Analysis, 159, 107217.

Zhu, C., Chang, Y., Li, X., & Shan, M. (2022). Factors influencing embodied carbon emissions of China's building sector: An analysis based on extended STIRPAT modeling. Energy and Buildings, 255, 111607.

Appendix

Appendix A- Hyperparameters of algorithms

Table A1. Data-splitting strategy and hyperparameter settings for the Boosting Regression model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training and validation data	80% of the full dataset
	Validation strategy	Random sampling
	K-fold cross-validation	Not applied
	Indicator variables	Not generated
Training parameters	Shrinkage (learning rate)	0.1
	Interaction depth	1
	Minimum observations per node	10
	Training data used per tree	50%
	Loss function	Gaussian
	Feature scaling	Applied
Ensemble configuration	Number of trees	Optimized
	Maximum number of trees	100
Reproducibility	Random seed	Not fixed

Table A2. Data-splitting strategy and hyperparameter settings for the Decision Tree Regression model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training and validation data	80% of the full dataset
	Validation sample	20% of training data
	K-fold cross-validation	Not applied
	Test set indicator	None
Algorithmic settings	Minimum observations for split	20
	Minimum observations in the terminal node	7
	Maximum interaction depth	30
	Feature scaling	Applied
Tree complexity	Complexity penalty	Optimised
	Maximum complexity penalty	1
Reproducibility	Random seed	Not fixed

Table A3. Data-splitting strategy and hyperparameter settings for the K-Nearest Neighbours (KNN) regression model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training and validation data	80% of the full dataset
	Validation sample	20% of training data
	K-fold cross-validation	Not applied
	Leave-one-out validation	Not applied
	Test set indicator	None
Algorithmic settings	Distance metric	Euclidean
	Weighting scheme	Rectangular (uniform weights)
	Feature scaling	Applied
Model complexity	Number of nearest neighbours (k)	Optimized
	Maximum number of neighbours	10
Reproducibility	Random seed	Not fixed

Table A4. Data-splitting strategy and parameter settings for the Linear Regression model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training data	80% of the full dataset
	Test set indicator	None
Algorithmic settings	Intercept term	Included
	Feature scaling	Applied
Reproducibility	Random seed	Not fixed

Table A5. Data-splitting strategy and hyperparameter settings for the Random Forest regression model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training and validation data	80% of the full dataset
	Validation sample	20% of training data
	Test set indicator	None
Algorithmic settings	Training data used per tree	50%
	Features per split	Auto
	Feature scaling	Applied
Ensemble configuration	Number of trees	Optimized
	Maximum number of trees	100
Reproducibility	Random seed	Not fixed

Table A6. Data-splitting strategy and hyperparameter settings for the Regularized Linear Regression (Lasso) model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training and validation data	80% of the full dataset
	Validation sample	20% of training data
	Test set indicator	None
Regularization settings	Penalty type	Lasso (L1 regularization)
	Regularization parameter (λ)	Optimized
	Largest λ within 1 SE of minimum	Not selected
Algorithmic settings	Intercept term	Included
	Feature scaling	Applied
Reproducibility	Random seed	Not fixed

Table A7. Data-splitting strategy and hyperparameter settings for the Support Vector Machine (SVM) regression model

Category	Parameter	Specification
Data split	Holdout test data	20% of the full dataset
	Training and validation data	80% of the full dataset
	Validation sample	20% of training data
	Test set indicator	None
Kernel and weighting	Weighting scheme	Linear
Algorithmic settings	Degree	3
	Gamma parameter	1
	r parameter	0
	Epsilon (ε)	0.01
	Termination tolerance	0.001
	Feature scaling	Applied
Regularization (C)	Cost of constraint violation (C)	Optimized
	Maximum violation cost	5
Reproducibility	Random seed	Not fixed

Appendix B- Summary Statistics

Table B1. Descriptive Statistics and Distributional Properties of Variables Used in the Empirical Analysis

	CBE	ASND	AGRL	AFFV	AFWW	CO2P	CDD	ECL	EIMP	ENIN	ENUP
Valid	513	443	487	513	461	435	435	502	486	487	509
Missing	0	70	26	0	52	78	78	11	27	26	4
Mode	0.096	0.005	16.883	3.083	21.546	7.642	2.377.810	0.000	-55.217	3.470	2.162.853
Median	5.706	0.099	45.603	1.998	15.040	6.755	294.510	16.023	57.621	3.470	2.918.138
Mean	19.802	0.241	41.534	2.252	21.667	7.307	528.338	22.322	65.823	3.633	3.340.367
Std. Error of Mean	1.411	0.016	0.719	0.056	1.138	0.161	27.614	1.046	2.557	0.051	60.065
95% CI Mean Upper	22.574	0.273	42.946	2.361	23.904	7.624	582.612	24.376	70.847	3.732	3.458.373
95% CI Mean Lower	17.030	0.210	40.122	2.142	19.430	6.990	474.063	20.268	60.798	3.533	3.222.362
Std. Deviation	31.957	0.336	15.859	1.262	24.443	3.363	575.946	23.428	56.376	1.119	1.355.117
95% CI Std. Dev. Upper	34.042	0.360	16.923	1.344	26.132	3.603	616.994	24.974	60.162	1.194	1.443.911
95% CI Std. Dev. Lower	30.114	0.315	14.921	1.189	22.960	3.153	540.048	22.062	53.040	1.053	1.276.673
MAD	4.636	0.084	11.103	0.916	10.102	1.823	230.380	14.299	19.568	0.640	755.285
IQR	21.882	0.270	21.969	1.808	19.333	3.600	619.850	30.543	38.970	1.405	1.625.971
95% CI Variance Upper	1.158.847	0.129	286.378	1.807	682.859	12.980	380.681.877	623.701	3.619.488	1.425	$2.085\times10^{+6}$
95% CI Variance Lower	906.826	0.099	222.649	1.414	527.178	9.944	291.651.794	486.754	2.813.291	1.108	$1.630\times10^{+6}$
Skewness	2.412	2.151	-0.571	0.934	1.850	2.116	1.752	1.298	4.217	0.556	1.440
Std. Error of Skewness	0.108	0.116	0.111	0.108	0.114	0.117	0.117	0.109	0.111	0.111	0.108
Kurtosis	5.637	4.548	-0.410	1.736	2.649	7.024	2.780	1.163	22.993	0.376	2.608
Std. Error of Kurtosis	0.215	0.231	0.221	0.215	0.227	0.234	0.234	0.218	0.221	0.221	0.216
Shapiro-Wilk	0.633	0.704	0.946	0.948	0.742	0.829	0.784	0.849	0.589	0.974	0.888
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
Range	168.962	1.840	60.396	8.482	104.705	22.635	2.377.810	95.710	505.398	6.370	8.118.257
Minimum	0.096	0.001	7.354	0.004	0.995	2.975	0.000	0.000	-55.217	0.970	1.311.667
Maximum	169.059	1.842	67.750	8.486	105.700	25.610	2.377.810	95.710	450.181	7.340	9.429.925
25th percentile	1.656	0.028	30.574	1.276	4.838	4.984	141.080	2.621	41.021	2.875	2.376.275
50th percentile	5.706	0.099	45.603	1.998	15.040	6.755	294.510	16.023	57.621	3.470	2.918.138
75th percentile	23.538	0.297	52.543	3.083	24.171	8.583	760.930	33.165	79.991	4.280	4.002.246
Sum	10.158.248	106.917	20.227.052	1.155.259	9.988.388	3.178.716	229.826.850	11.205.577	31.989.860	1.769.070	$1.700\times10^{+6}$

Note: The table reports descriptive statistics for all variables used in the analysis, including central tendency, dispersion, distributional shape, and normality tests. Missing values reflect data availability across countries and years in the unbalanced panel.

Table B1 (Cont.). Descriptive Statistics and Distributional Properties of Variables Used in the Empirical Analysis

	FPI	FRST	FOSC	HDD	WSTR	CH4P	N2OP	PM25	RELE	RENC	TCL
Valid	487	487	509	435	461	435	435	435	461	461	444
Missing	26	26	4	78	52	78	78	78	52	52	69
Mode	141.110	18.626	97.520	1.309.940	22.972	0.662	0.379	19.706	0.019	3.300	0.000
Median	98.010	34.058	73.460	5.526.790	17.974	0.899	0.527	14.850	25.112	16.400	13.006.500
Mean	98.785	34.507	73.063	5.488.345	21.403	1.011	0.675	15.124	30.005	18.699	39.071.032
Std. Error of Mean	0.558	0.766	0.706	100.052	0.916	0.025	0.021	0.246	1.033	0.566	3.127.076
95% CI Mean Upper	99.882	36.013	74.449	5.684.991	23.203	1.059	0.716	15.607	32.036	19.812	45.216.779
95% CI Mean Lower	97.688	33.001	71.676	5.291.699	19.604	0.963	0.634	14.642	27.975	17.586	32.925.284
Std. Deviation	12.324	16.915	15.922	2.086.743	19.662	0.513	0.435	5.122	22.188	12.160	65.891.590
95% CI Std. Dev. Upper	13.151	18.050	16.965	2.235.469	21.020	0.549	0.466	5.488	23.721	13.001	70.536.492
95% CI Std. Dev. Lower	11.596	15.915	15.000	1.956.680	18.469	0.481	0.408	4.803	20.842	11.423	61.823.992
MAD	4.860	5.946	9.670	1.044.180	11.910	0.137	0.124	4.066	14.690	7.800	11.694.500
IQR	9.710	17.100	19.470	2.090.220	23.765	0.270	0.294	8.223	32.159	17.400	38.258.000
95% CI Variance Upper	172.957	325.796	287.814	$4.997\times10^{+6}$	441.842	0.302	0.217	30.113	562.668	169.016	$4.975\times10^{+9}$
95% CI Variance Lower	134.468	253.295	225.004	$3.829\times10^{+6}$	341.109	0.231	0.167	23.070	434.390	130.483	$3.822\times10^{+9}$
Skewness	1.240	0.466	-0.727	-0.096	1.352	3.579	1.799	0.111	0.712	0.801	2.722
Std. Error of Skewness	0.111	0.111	0.108	0.117	0.114	0.117	0.117	0.117	0.114	0.114	0.116
Kurtosis	3.735	0.071	0.600	0.278	1.607	14.279	3.000	-0.733	-0.395	0.046	7.316
Std. Error of Kurtosis	0.221	0.221	0.216	0.234	0.227	0.234	0.234	0.234	0.227	0.227	0.231
Shapiro-Wilk	0.884	0.942	0.958	0.975	0.859	0.592	0.790	0.984	0.933	0.940	0.602
P-value of Shapiro-Wilk	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
Range	75.420	72.642	75.660	10.539.110	90.294	3.435	2.255	22.867	89.041	57.800	339.968.000
Minimum	65.690	1.094	24.280	743.700	0.993	0.440	0.093	4.895	0.003	0.100	0.000
Maximum	141.110	73.736	99.940	11.282.810	91.287	3.875	2.348	27.762	89.045	57.900	339.968.000
25th percentile	92.865	22.726	65.140	4.502.860	5.999	0.795	0.420	11.134	12.342	9.200	4.329.750
50th percentile	98.010	34.058	73.460	5.526.790	17.974	0.899	0.527	14.850	25.112	16.400	13.006.500
75th percentile	102.575	39.826	84.610	6.593.080	29.764	1.065	0.715	19.357	44.501	26.600	42.587.750
Sum	48.108.370	16.805.055	37.188.980	$2.387\times10^{+6}$	9.866.916	439.791	293.443	6.579.089	13.832.439	8.620.200	$1.735\times10^{+7}$

Note: The table reports descriptive statistics for all variables used in the analysis, including central tendency, dispersion, distributional shape, and normality tests. Missing values reflect data availability across countries and years in the unbalanced panel.

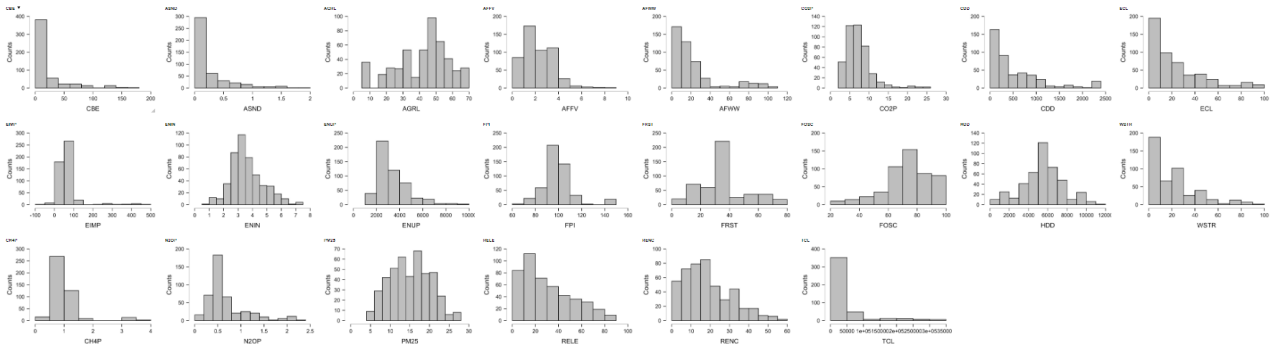

Figure B1. Distribution of Key Variables Used in the Empirical Analysis (Note. The figure displays histograms for all variables included in the study, illustrating their empirical distributions, skewness, and dispersion. The plots highlight non-normality and heterogeneity, supporting the use of panel estimators robust to distributional deviations.)

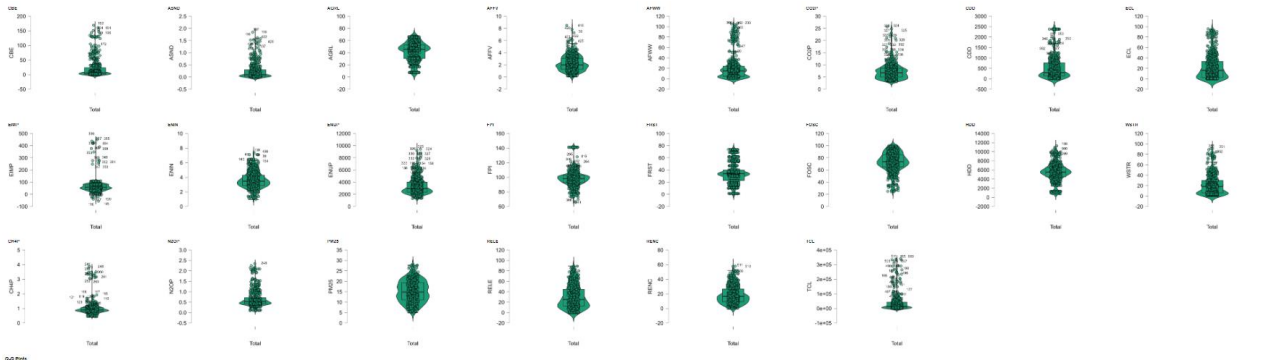

Figure B2. Violin Plots of Key Variables in the EU Building-Sector Emissions Dataset (Note: The figure presents violin plots illustrating the distribution, central tendency, and density of all variables used in the analysis. It highlights dispersion, asymmetry, and outliers, reinforcing the presence of heterogeneity across countries and time periods.)

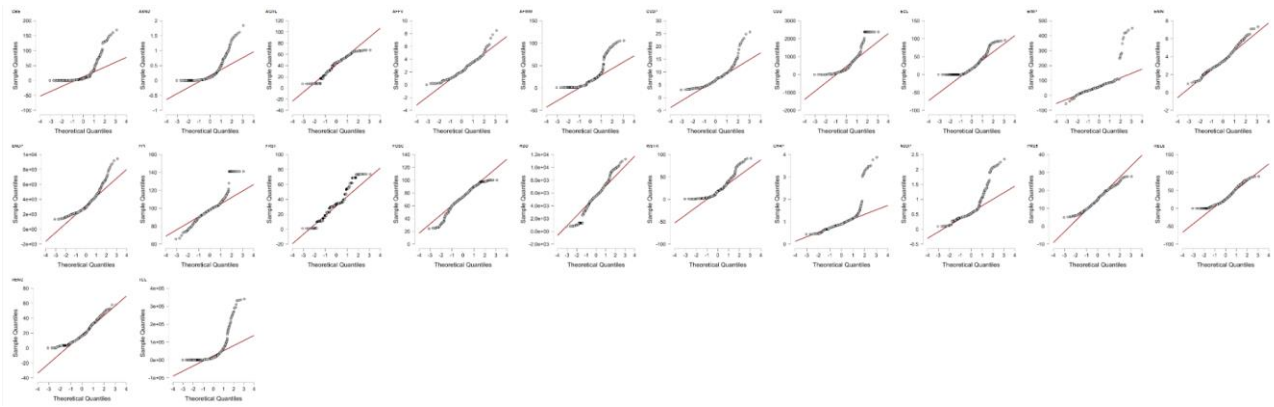

Figure B3. Quantile–Quantile (Q–Q) Plots for Normality Assessment of Model Variables (Note: The figure displays Q–Q plots comparing empirical distributions of each variable with a normal distribution. Deviations from the reference line indicate skewness, heavy tails, and non-normality, supporting the use of robust panel estimators and alternative modeling strategies.)