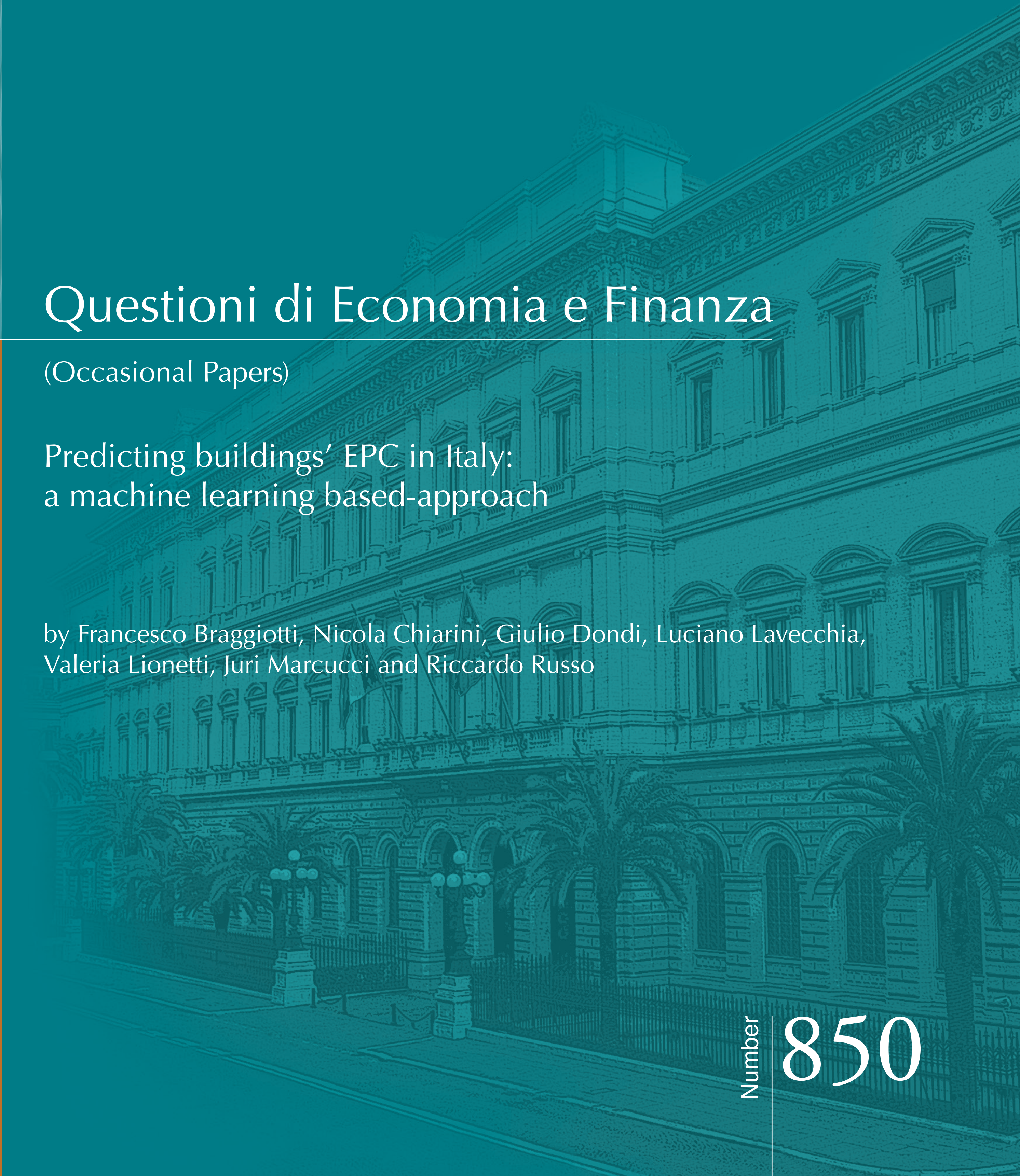
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

Predicting buildings' EPC in Italy: a machine learning based-approach

by Francesco Braggiotti, Nicola Chiarini, Giulio Dondi, Luciano Lavecchia, Valeria Lionetti, Juri Marcucci and Riccardo Russo

June 2024

Number 850

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Predicting buildings' EPC in Italy: a machine learning based-approach

by Francesco Braggiotti, Nicola Chiarini, Giulio Dondi, Luciano Lavecchia, Valeria Lionetti, Juri Marcucci and Riccardo Russo

Number 850 – June 2024

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of the Bank of Italy

PREDICTING BUILDINGS' EPC IN ITALY: A MACHINE LEARNING BASED-APPROACH

by Francesco Braggiotti*, Nicola Chiarini*, Giulio Dondi*, Luciano Lavecchia**,
Valeria Lionetti**, Juri Marcucci** and Riccardo Russo**

Abstract

EU member states have committed to achieving carbon neutrality by 2050. Given that building-related activities contribute to almost a quarter of EU greenhouse gas emissions, the implementation of new regulations to decarbonize this sector is paramount. However, policymakers must carefully evaluate policies to mitigate transition risks associated with these regulations, as buildings represent a significant portion of household wealth and bank assets. Accurate metrics regarding buildings' energy efficiency, including the energy class reported in Energy Performance Certificates (EPCs), are essential for such evaluations. In this study, we developed a machine learning-based model to predict the energy class of Italian buildings using publicly accessible data. The model, trained on a geographic subset of the Italian territory, achieves a 37% accuracy rate, which increases to 74% when allowing for a one-class margin of error (1-class accuracy). Further testing against a mortgage portfolio provided by a commercial bank yielded a 69% 1-class accuracy. Comparison with statistics reported by the official EPC register (SIAPE) suggests a potential discrepancy in the representation of the worst energy efficiency class.

JEL Classification: Q47, Q54, Q58, C45, C53, C55.

Keywords: energy performance certificates, EPC, Italy, buildings, transition risk, machine learning, random forest classifier, random forest.

DOI: 10.32057/0.QEF.2024.0850

* Datasinc.

** Bank of Italy, DG Economics, statistics and research (Marcucci and Russo) and Climate change and sustainability hub (Lavecchia and Lionetti).

1 Introduction[1]

Decarbonizing the energy use of the building sector is fundamental for mitigating climate change and achieving the ambitious climate targets engraved in the Paris Agreement. Buildings consume a significant amount of energy (approximately 35% of the final energy consumption in Europe[2]), and were responsible for almost a quarter of EU GHG emissions in 2021.[3] However, energy retrofits are costly and lengthy.

At the same time, real estate is a significant part of both households' wealth and banks' assets. Real assets, such as houses, account for a significant portion of households' gross wealth (82% as of the end of 2020 in Italy; Banca d'Italia 2022). Residential real estate (RRE) and commercial real estate (CRE) loans are a significant part of banks' portfolios while mortgages are usually collateralized and bought by insurance and pension funds because of their relatively stable cash flows and low risk.

Some jurisdictions, like England and France, have already implemented mandatory regulations to enhance the energy efficiency of houses. For instance, since April 1, 2018, a Minimum Energy Efficiency Standard (MEES) has been enforced in England and Wales. This regulation prohibits the rental or rent-to-buy of any house with an Energy Performance Certificate (EPC) rating below class E. Similarly, in France, as of January 1, 2025, a house with an EPC[4] rating below F will no longer be available for rent.[5] Such measures represent significant changes, potentially impacting owners' real wealth.

Meanwhile, the European Union, known for its staunch stance against climate change,[6] has recently established new targets through the recasted Energy Per-

[1] The views expressed are those of the authors and do not necessarily reflect those of the institutions to which they belong. This paper is the result of the collaboration between the privately-owned company Datasinc and the Bank of Italy, within the first Call for Proposals by the Bank of Italy's Milano Hub. The data utilized in this study were collected by Datasinc, a company acting under its sole responsibility. They were graciously provided to researchers at the Bank of Italy in an aggregated format. Neither the Bank of Italy nor its employees were involved in the data collection process or contributed to the creation of the dataset in any way. The aggregate data shared with the Bank of Italy were solely used by the institution for research purposes. We would like to thank Paolo Angelini, Diletta Arcangeletti, Laura Bartiloro, Giuseppe Bruno, Luigi Cannari, Ivan Faiella, Alberto Felettigh, Massimo La Rana, Stefano Montemaggi, Patrizio Pagano, Valeria Passano, Marino Ottavio Perassi, Alessandra Perrazzelli, Alfonso Rosolia, Paolo Sestito, Bruna Szego, Alberto Ventura and Ylenia Yashar for their useful comments.

[2] Energy efficiency trends in buildings in the EU, ODYSSE-MURE.

[3] Greenhouse gas emissions from energy use in buildings in Europe, European Environment Agency.

[4] EPCs across countries are not comparable.

[5] Interdiction de location et gel des loyers des passoires énergétiques, Ministeŕe de la Transition Ecologique et de la Cohes̀ion des Territoires.

[6] The European Climate law includes ambitious goals such as achieving carbon neutrality by 2050 and reducing greenhouse gas emissions by at least 55% by 2030, compared to 1990 levels.

formance of Buildings Directive (EPBD).[7]

Policies mandating energy retrofits could influence the pricing of this crucial asset class. Therefore, a careful evaluation before implementation is essential. This paper contributes to the literature on real estate transition risk (Bernardini et al., 2021) by presenting a machine learning approach for predicting the Energy Performance Certificates (EPCs), known in Italian as *Attestato di Prestazione Energetica*, for all buildings in Italy.

In Italy, every building sold, rented, or undergoing major renovation must obtain an Energy Performance Certificate (EPC), which is then transmitted to the regional energy cadastre. Currently, there are over 67 million building units (of which 36 million are houses) in Italy, yet as of May 2024 only 5.7 million EPCs are registered in Italy's national EPC cadastre, known as the *"Sistema Informativo sugli Attestati di Prestazione Energetica"* (SIAPE).

The SIAPE register, managed by ENEA (Italy's national energy efficiency agency), was established in 2015 and has been gradually populated since then for all italian regions (except for Campania and Sardinia). The coverage of this database has only recently become substantial. However, over 1 million EPCs failed a quality check conducted by ENEA and have been filtered out, and some regions have only partially transmitted their archives; for example, while there are 223 thousand EPCs from Sicily in SIAPE, the regional cadastre contains over 690 thousand units as of May 2024.

As a result, the majority of buildings in Italy lack any assessment of their energy performance. Large European banks have only recently begun collecting EPCs due to Pillar 3 ESG disclosure obligations. Consequently, this information is absent from the majority of mortgage portfolios.

Our study is the first attempt to estimate the energy efficiency class of each building in Italy. This information is crucial for evaluating the transition risk associated with a mortgage portfolio or any derivatives.

Energy certificates are issued by an expert following an on-site visit, during which critical aspects such as the building's construction, fixtures, heating type, floor, position etc., are evaluated. Similarly, our model predicts the energy efficiency of Italian buildings using cadastral data and other open data sources, including building/apartment features, municipality characteristics, etc. We trained our model on data from the Lombardy region.[8]

[7] The directive sets objectives for reducing the average primary energy consumption for buildings; establishes criteria for the national building renovation plan; requires Member States to promote the introduction of investment and financing instruments, such as, for example, energy efficiency loans and green mortgage for the renovation of buildings in addition to public finance interventions.

[8] EPCs are collected by each region. However, only Lombardy, Piedmont and Trentino published their archives as open data. In this work we use mainly data from Lombardy (the most populated region in Italy) and those of Piedmont as a robustness check.

Our model accurately predicted the energy class in 37% of cases. This rate increases to 74% when allowing a one-class margin of error. In the case of Piedmont, used as robustness, the respective figures are 25% and 61%.

We tested our model in a real-world scenario through a collaboration with an Italian bank. Specifically, we applied our model to estimate the energy class of a set of properties serving as collateral for a mortgage portfolio in the Lombardy region. Among the real estate units in this portfolio, our model accurately predicted the energy class in 31% of cases (69% when allowing for a one-class margin of error).

While these performance levels are lower compared to similar attempts documented in the literature, such as deep learning models coupled with street view imagery (Sun et al., 2022), which require advanced IT infrastructures and are not freely accessible, our model's performance remains substantial. Furthermore, our model's performance aligns closely with another recent study conducted in the UK (Mayer et al., 2023) albeit with a more ambitious aim of class-by-class prediction, as opposed the UK study which focused on larger aggregates (class A-D vs E-G). Finally, our model is parsimonious with data, requiring only a dozen variables.

We also compared the distribution of the energy classes across all buildings in Italy from our model with the distribution from the national cadastre (SIAPE). We identify two significant differences: our model underestimates the proportion of buildings in all classes except for class F (one of the least efficient), consequently overestimating the proportion of buildings in class F (in the case of residential buildings, by 35 percentage points: 59% in our model vs. 24% in SIAPE). This discrepancy likely arises from a sample selection bias in the SIAPE dataset, as the issuance of EPCs is legally required for new constructions, renovations, or energy retrofits— situations where buildings are more likely to be efficient. As a result, the distribution of buildings in SIAPE may be distorted towards more energy-efficient houses compared to the actual stock of all buildings. Based on the lower end of the energy efficiency distribution (classes G and F), our estimates suggest that 82% of Italian houses are inefficient, contrasting with the 56% inferred from the SIAPE distribution. This disparity holds significant implications for policy formulation, highlighting the necessity for energy retrofitting in approximately 9 million additional houses.

2 Literature review

We have split our literature review into two sections. The first section delves into regulatory and supervisory aspects, while the second section explores the technical literature on research close/similar to ours.

2.1 EU bank regulatory and supervision framework

Climate change affects the banking system through the impact on macro and microeconomic variables, generating different types of risk that influence the typical risks of banking institutions. The transition to a carbon-free economy indeed generates transition risks due to changes in government policies, technological developments, or changes in consumer and investor preferences. Examples of such changes are the introduction of energy efficiency policies or taxes on fossil fuels or incentives for the use of eco-sustainable sources; other examples include the introduction of technologies with a lower environmental impact or re-direction of the choices of consumers and investors towards more eco-sustainable products and services. At an international and European level, standard setters and regulators (BCBS, 2021; NGFS, 2020; EBA, 2021) agree on conducting analyses of the impact on banks' balance sheets that these risk factors have on the existing traditional categories of financial risks (credit, market, liquidity, operational and reputational). For example, a borrower's ability to repay his debt could be affected by extreme weather events in some geographic areas or customers' changing sensitivity to climate-related issues could damage the bank's reputation when lending to sectors of controversial activities on environmental issues. Real estate assets used as collateral in banks' credit portofolio, with the aim of credit risk mitigation, are also exposed to transition risks. Energy consumption and prices indeed have an impact on the value of properties and therefore on their ability to mitigate the credit risk associated with secured exposures. Changes in regulatory standards on the energy efficiency of buildings[9] could lead to a deterioration in the value of those less energy efficient and make it more difficult to sell them for the recovery of the credit guaranteed by these properties. The transition risk could also affect the value of properties due to the change in consumer preferences (RICS, 2022) or changes in the demand and prices of energy-efficient properties especially in the context of rising energy prices such as those we've been experiencing since mid-2021. A recent study on the Belgian real estate market (Reusens et al., 2022) shows that in the last ten years, the price difference between energy-efficient and energy-guzzling houses is large and has been increasing. In 2021 a house in Belgium with energy label B is about 12% more expensive than a similar house with energy label D and 22% more expensive than a similar house with energy label F. Also in Italy,[10] recent research finds that the best energy-performing houses sell at a 25% premium compared to the worst-performing ones (Loberto et al., 2023). Concerning the im-

[9] A prescription similar to the English minimum energy efficiency standard - mentioned in the introduction - has been in force in the Netherlands for corporate real estate since 1 January 2023.

[10] The energy class identifies the energy performance of a building, i.e. the amount of energy required to meet the requirements related to the standard use of a building, annually, for heating, cooling, ventilation, production of domestic hot water and, in non-residential buildings, lighting, elevators, and escalators. There are ten energy classes, ranging from A4 for the most efficient buildings to G for the least efficient. This classification was introduced in Italy by Decreto interministeriale 162/15 following the transposition of the European Directive 2012/27. The energy performance certificate (EPC) is known in Italian as *"Attestato di Prestazione energetica"* or APE.

pact of transition risk and creditworthiness, numerous studies have analyzed the relationship between the probability of default of mortgages and the energy efficiency of the properties used as collateral. A study by the Bank of England[11] has highlighted how the energy efficiency of a building could be a relevant predictor of mortgage defaults. A recent study by Energy Efficiency Financial Institutions Group[12] concludes that there is a statistically significant correlation between the energy performance of building collateral and mortgage credit performance.

In light of their potential impact on the financial stability of supervised institutions and the whole system, the inclusion of climate-related and environmental risks in the prudential regulatory framework and the supervisory ongoing activities is a priority in the agenda of international standard setters, regulators, and supervisors. In the European jurisdiction, the European Commission published in 2018 the "Action Plan on financing sustainable growth" (European Commission, 2018) with the aim – through even changes in the financial regulatory framework - of redirecting capital flows towards a more sustainable economy; managing financial risks arising from climate change, resource depletion, environmental degradation and social issues; promote transparency and long-term vision in economic and financial activities. In 2020, the European Commission also adopted the European Green Deal (European Commission, 2019) to reduce net greenhouse gas emissions by at least 55% by 2030 and to reach climate neutrality by 2050. To realign capital flows towards meeting these objectives and mitigate the risk of greenwashing practices, the European Commission introduced the Taxonomy Regulation (Commission, 2020), aimed at establishing criteria for classifying economic activities as sustainable, thereby aligning related financial assets with the established taxonomy. Taxonomy Regulation applies to financial institutions that offer financial products on the European market and also includes (article 8) disclosure requirements for financial and non-financial companies already under the Corporate Sustainability Reporting Directive.[13] Since 2004, disclosure requirements have included the *Green Asset Ratio* (GAR), which is the proportion of company activities aligned with the taxonomy. For credit institutions, the GAR represents the proportion of assets financed and invested in taxonomy-aligned economic activities, as a percentage of

[11] Guin and Korhonen (2020) assessed a sample of 1.8 million outstanding mortgages throughout the UK and found that energy-efficient mortgages (rated A–C) have 11 bps lower payments arrears shares than low-energy-efficiency mortgages (rated D–F). This result is more robust after controlling for factors such as borrowers' income, borrowers' age, LTVs, property, and regional factors.

[12] European Commission and Directorate-General for Energy (2022) considered a sample of about 800,000 residential mortgages granted in the UK, Germany, Finland, and Italy and included also a review of the literature on green premiums in real estate in several European countries. The Energy Efficiency Financial Institutions Group (EEFIG) is a specialist expert working group by the European Commission and United Nations Environment Programme Finance Initiative (UNEP FI) and also includes representatives of the financial and banking sector, representatives of industry, and energy efficiency specialists.

[13] The Corporate Sustainability Reporting Directive (CSRD) adopted in 2022 applies to companies with more than 250 employees and listed SMEs. A detailed overview of ESG disclosure requirements for banks can be found in Loizzo and Schimperna (2022).

total covered assets (including the percentage of financial guarantees supporting debt instruments financing taxonomy-aligned economic activities). In 2021, the EU Commission adopted the first Technical Screening Criteria (TSC) relating the climate change adaptation and mitigation objectives.[14] The first TSC includes criteria for buildings classification[15] when they are used as collateral in a secured portfolio, with a significant impact for banks to determine taxonomy compliance of the underlying mortgage and loans (as well as a covered bond for example). The energy efficiency rating or score of a building is the criteria used in the TSC, as a proxy of the contribution of the building to taxonomy objectives.[16]

The EU Commission's Action Plan also steered the European Banking Authority's (EBA) Sustainable Finance Action Plan in 2019 which progressively introduced ESG factors into the three pillars of prudential regulation for European banks. In June 2021, EBA published the "Report for the management and supervision of ESG risks for banks and investment firms" (EBA, 2021), focused on assessing the resilience of banks and investment firms to the potential impacts of ESG risks across different time horizons, and recommends to institutions and supervisors a holistic, forward-looking and proactive approach on ESG risks. In February 2022, the EBA published an implementing technical standard (ITS) on prudential disclosure of ESG risk to broaden the Pillar 3 framework and enforce market transparency (EBA, 2022). The technical standard applies to listed large institutions[17] and requires the publication, from 2023, of qualitative and quantitative information on ESG risks, with a specific focus on climate risks, including GAR and Banking Book Taxonomy Alignment Ratio (BTAR). In template 2 of the ITS, banks should disclose information on loans collateralized by immovable property, detailing energy consumption and the energy class of the collateral. In October 2023, the EBA published a report (EBA, 2023) to assess how the current prudential framework captures environmental (and social) risks and recommends targeted enhancements to accelerate the integration of these risks across Pillar 1. The report encourages the inclusion of energy efficiency to address transition risk in the section on due diligence requirements and valuation of immovable property collateral.

[14] In the Taxonomy Regulation, an activity is deemed as sustainable if it: i) substantially contributes to one or more of the six environmental objectives; ii) does not significantly harm (DNSH) to the other objectives, and iii) meets minimum social safeguard standards; iv) complies with technical standard criteria that have been established by the Commission. The six environmental objectives identified by the legislator are i) climate change mitigation; ii) climate change adaptation; iii) sustainable use and protection of water and marine resources; iv) transition to a circular economy; v) pollution prevention and control; and vi) protection and restoration of biodiversity and ecosystems. Currently, there are Technical Screening Criteria only related to the first two objectives.

[15] The classification regards: i) construction of new buildings; ii) renovation of existing buildings; and iii) acquisition and ownership of buildings.

[16] In the first TSC, the alignment to adaptation and mitigation criteria considers buildings with an energy class "A" or, as an alternative, the 15% best in the class of the regional and national building stock.

[17] In the CRR3 proposals (published by the European Commission in 2021), the application of this ITS could be extended to all European banks (Commission, 2021).

Supervision authorities may contribute to raising awareness of the impact of climate and environmental risks on banks' activities and stimulate their inclusion in banks' practices and policies by setting their expectations (NGFS, 2020). In November 2020, the European Central Bank published a "Guide on climate-related and environmental risks" for Significant Institutions (ECB, 2020), setting out 13 supervisory expectations regarding how climate and environmental risk could be integrated into the strategy, business model, governance, risk management framework, and disclosure under the current regulatory framework. The Guide aims to promote sound and effective practices in climate and environmental risk management and disclosure and to provide a starting point for the supervisory dialogue.[18] In expectation 8 of the Guide ("Credit risk management"), the ECB explicitly refers to the holistic approach in the integration of climate and environmental risks at all stages of the credit process as envisaged in the EBA Guidelines on Loan Origination and Monitoring (LOM). These guidelines, published in June 2020 (EBA, 2020), recommend the inclusion of ESG factors in credit policies and procedures, in creditworthiness assessment and collateral. In collateral management, the ECB Guide requires that the impact of climate and environmental risks in the collateral assessment considers the physical location and energy efficiency of residential and non-residential buildings. These aspects should be integrated both in the process of establishing the value of the collateral and in the review process provided for by the applicable regulation. The policy on credit mitigation tools should be reviewed to take into account the assessment of these climate-related metrics. The EBA LOM also recommends evaluating the use of energy efficiency performance and EPC labels for the assessment of residential and commercial property value.

2.2 Modeling the energy performance of buildings

Modeling energy performances is paramount to enhancing energy efficiency and reducing GHG emissions. However, lack of data, and low computational power, have historically hampered most of the attempts to model energy demand. More recently, large and open datasets, joined with new techniques based on advanced statistics and machine learning are providing a rapid increase in model availability.

A way to improve buildings' energy efficiency is to simulate their performance under different circumstances in the design phase - modeling energy demand before construction is fundamental, with significant gains (up to 40%) just simply getting the proper shape and orientation (Sahu et al., 2012). As a way to reduce the costs of multiple designs, Elbeltagi and Wefki (2021) propose an artificial neural network (ANN) algorithm to help architects reduce simulation costs and find the best solution to enhance energy performances. Li and Yao (2021) developed a framework to predict the energy consumption of residential and non-residential buildings by generating energy consumption data feeding different machine learning models.

[18] In April 2022, the Bank of Italy also published supervisory expectations for all supervised entities on how to integrate climate and environmental risk in their current practices.

In the United Kingdom, there is a long history of developing methodologies to assess and compare the energy and environmental performance of dwellings. Indeed, the Building Research Establishment (BRE), a profit-for-purpose organization, has been feeding HM Government with several tools since 1992 to calculate the energy use and fuel requirements of dwellings, based on their characteristics. The most important is the so-called BRE Domestic Energy Model (BREDEM), whose latest version is the BREDEM 2012 (updated in 2015) and the standard assessment procedure (SAP) methodology. The latter has been used since 1994 to assess the energy performance of new buildings, and, lately, to produce the Energy Performance Certificate (EPC). In 2005 a Reduced Data SAP (RDSAP) was introduced as a simplified method of assessing the energy performance of existing dwellings. The current versions are SAP (10.2) and RDSAP 2012. Both assess how much energy a dwelling will consume given a defined level of comfort and service provision, an energy efficiency rating, and the CO2 emissions. All these models are trained and calibrated using the yearly English housing stock (EHS), a survey on dwelling energy demand and performance.

A recent study (Williams and Bonham, 2020) conducted by the Data Science Campus organization, sponsored by the Welsh Government, developed a predictive model for rating the energy efficiency of homes in Wales, as part of an initiative to reach zero-carbon emissions by 2050. The purpose was to apply this model to all buildings for which the EPC has never been produced, thus obtaining a comprehensive, nationwide picture of energy efficiency in the residential sector. Similarly to Italy, most buildings in the UK have never undergone an evaluation procedure to measure energy scores. The model was implemented in a two-step procedure in which a prototype model was first obtained by training the eXtreme Gradient Boosting (XGboost) algorithm on an open dataset of 700k EPC-evaluated welsh buildings, attaining 82% accuracy. This first step provided insight into the most important features for determining EER scores, such as total area, floor height, and hot water system. The second step was aimed at finding proxy variables for these features, as the latter are only available for EPC-evaluated buildings (about 50% of the total). Many of these proxy variables were obtained with closest neighbors techniques in conjunction with the EPC-evaluated dataset. The proxy-based model attained a 40% accuracy, suggesting a significant loss of information when attempting to reconstruct all features annotated in EPC Certificates via an indirect approach.

Artificial neural networks (ANN) have also been employed for similar estimations in recent works. Lin et al. (2022) report the results of 24 energy consumption models trained on a sample of 227 houses in Oshawa (Canada). Data collected via smart metering, phone surveys and energy audit are used to train eight data-driven algorithms,[19] the most accurate being the Backpropagation Neural Network

[19] These are: Multiple Linear Regression (MLR), Stepwise Regression (SR), Support Vector Machine (SVM), Backpropagation Neural Network (BPNN), Radial Basis Function Neural Network

(BPNN).

In Sun et al. (2022), a DenseNet-based architecture has been trained on EPC data coupled to Google Street View (GSV) imagery, attaining an overall accuracy of 86%. Explainable AI techniques have shown that insulations around doors and windows strongly affect energy efficiency.

To further develop an EPC-independent ANN classifier Mayer et al. (2023) have trained a model on fully remotely sensed data sources (GSV and Landsat-8 satellite images). This model achieved a 68.30% accuracy in predicting two macro-classes, namely A-D and E-G energy rating intervals.

The approach described in the following sections largely resembles the one used in Williams and Bonham (2020), given the set of data sources currently at our disposal. Retrieval and inclusion of image data in the model is an interesting development that could be approached in future works.

3 Data

The ensuing sections encapsulate a comprehensive overview of the diverse data sources harnessed in the course of this investigation.

3.1 Real estate data

In 1861, with the unification of the Italian Crown, Italy had 22 different Cadastres, each with its logic and boundaries. The first law to unify all properties into a single Ordinary Cadastre was enacted in 1886. This law created a set of Sheets ("*Fogli*") for each Municipality ("*Comune*") upon which all land portions and buildings were drawn.[20] In turn, each land portion and/or building was identified as a parcel ("*Particella*" or "*Mappale*"). Furthermore, within each parcel, there is a sub-portion ("*subalterno*") code to identify a specific unit. Summing up, each house can be identified using a combination of Municipality ("*Comune*"), Sheet ("*Foglio*"), Parcel ("*Particella*"), and sub-portion ("*subalterno*").

This paper focuses on the Ordinary Cadastre, which covers more than 95% of all properties in Italy.

The Ordinary Cadastre is formed by two layers:

(RBFN), Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), and Exhaustive CHAID (ECHAID)

[20] While this was true for most regions, some of the North-Eastern provinces did not adopt the Italian Ordinary Cadastre but kept the Austrian Empire system, the so-called "*Catasto fondiario*" or "*Tavolare*". The main difference between the two is that the *Tavolare*' is probatory and only a judge can change the deeds, the Ordinary Cadastre is not. During the '900, most Italian regions adopted the Ordinary Cadastre except for Trentino Alto Adige, the provinces of Trieste and Gorizia, and some smaller municipalities in Lombardy and Veneto. As of today the two Cadastres still co-exist.

- **Land portions**: pieces of land used for agriculture, parks, etc.
- **Buildings**: once developed on a piece of land, buildings substitute the land portion, which becomes a sub-portion of the building. Other ancillary units can be sub-portioned to the main unit (e.g. garages, attics, etc.).

Theoretically, all land portions should correspond one-to-one with buildings (*"Particelle"*), but in practice, more than 20% of the time, this is not the case. This discrepancy arises because Municipalities have undergone split or merge processes over the years, resulting in new numbering for each Sheet and Parcel. Newly assigned numbers have been applied to all newly transacted buildings, while those that have never been transacted still retain old numbers. While land portions have been updated with new numbering, the old numbering for building portions has been retained, leading to duplicates that do not match in terms of either name or numbers, thus causing misalignment and confusion. This misalignment generates a critical issue as the *Agenzia delle Entrate* (the Italian Income Revenue Authority who manages the Cadastre) only provides mapping of the land portions. Moreover, neither the Sheets nor the Parcels numbers are sequential but placed semi-randomly on the map (e.g., Sheet 1 and Sheet 2 may or may not be next to each other and so are Parcels).

As this misalignment is structural in the Ordinary Cadastre, we used a proprietary tool, namely *Elenco Immobili* which can correctly map buildings on the land mapping.[21] All non-matching properties are re-mapped either through the (verified) street address or through an in-depth analysis of all related data available in the Cadastre.

Datasinc has developed a tool that periodically maps the entire Cadastre from the Sister website[22] tapping into each municipality, each "*foglio*", each "*Particella*" and each "*Subalterno*". In other terms, the machine gets all the data available on each page within the website for all the building units in each municipality and adds it to a Postgres relational database. The entire database, collecting more than 180 million units, represents a complete mapping of the Italian real estate stock (buildings and land portions) and is periodically updated to track changes in the Italian real estate environment.

[21] *Elenco Immobili* is a proprietary database owned and maintained by Datasinc, collecting Cadastre data for over 180 million data points. The database is created from the *Agenzia delle Entrate*'s portal, keeping track of all changes and updates at the building/unit level. The floor area of buildings is not published and cannot be retrieved via the aforementioned procedure. However, the portal offers access to the cadastral income ("*Rendita catastale*") values (in euros), which are strictly correlated to the floor surface.

[22] Sister is the service providing all cadastral information belonging to *Agenzia delle Entrate* for public use. It requires a registration and does not provide any API for faster interactions. Within the website, units or subjects can be searched with the right coding (VAT number for subjects or Sheet, Parcel for units). The response contains all the details on the building (e.g. address, tax value, size, etc.) and allows for a further search on all beneficiaries of the unit.

As of today, to summarize, the data collected in the *Elenco Immobili* dataset include, as of the time of writing:

- 180.45 million total units, of which 109.33 million are land portions, and 71.13 million are building portions.
- Within building portions, 55.3% are cadastral category A buildings (i.e., residential apartments, villas or offices), 41.3% are category C (i.e., non-revenue-generating units such as garages, warehouses, etc.), 2.8% are category D (i.e., revenue-generating units for industrial or commercial purposes) and the remaining categories sum up to 0.6%.
- Within building portions, 11.5% are in Lombardy, 10.0% in Sicily, 9.3% in Piedmont, 8.0% in Veneto.

For each of these units, the database includes:

- Type of property (land vs. buildings and the corresponding cadastral category)
- Municipality and Street address
- Floor
- cadastral income ("*Rendita catastale*"), assigned by the tax authority
- Dimension of the unit (in m^2)

Street addresses are of limited help. First of all, no land portion has a street address. This is because typically they are in rural areas or portions of land reachable only on unpaved roads. Moreover, even geo-referencing buildings using street addresses and available mapping tools on the web result in unsatisfactory precision. To be sure of the GPS coordinates, Datasinc developed, in collaboration with the International School for Advanced Studies (SISSA) in Trieste, a tool that extracts and georeferences polygons for each entity in the Italian Cadastre. These polygons are computed by scanning raster maps from the Cadastral databases, and their centroid can be used for localizing each building with an approximate error of ±9 meters, which is negligible for the scope of our analysis.

We use the information on the floor to create a dummy variable (*boundary unit*, see Table 3) which is equal to 1 whenever the unit is located either on the ground or the highest floors, i.e. where thermal dispersion is higher.

3.2 Energy performance certificates

According to current Italian legislation ("*Decreto Interministeriale* 26/06/2015") EPCs provide the energy class, EPGL ratings,[23] and other information such as

[23] EPGL is an index describing the overall energy consumption per square meter for heating, cooling, lighting, ventilation, and hot water production. EPGL.nren defines the same consumption in a

heated (cooled) surface, floor per unit, construction year and climate zone. They should be collected by each region though only Lombardy, Piedmont, and Trentino have published their archives as open data, with Lombardy's archive (CENED) being the most extensive. The Italian Energy Efficiency Agency (ENEA) owns the *SIAPE* database, which collects most of the energy certificates at the national level. Unfortunately, the data are currently published on an aggregated basis and do not cover the entire Italian territory.

We trained our algorithm on a subset of the Lombardy (CENED) database, specifically focusing on the EPCs issued after 2015 (approximately 1.21 million EPCs at the time of the analysis).[24] Some of these certificates have been excluded from the dataset due to errors; for instance, a significant number of certificates refer to units that should not have an energy certificate (such as land portions, garages, etc.). Furthermore, to establish a benchmark by region/province, the distribution of energy efficiency classes by province has been retrieved from the SIAPE website (which only publishes aggregated data).

3.3 Other datasets

An energy class prediction model is valuable for making predictions for buildings without certification. Therefore, CENED data have only been utilized to extract the target variable, namely the energy class, while various strategies have been applied to generate proxy variables from different sources, as listed below. The following paragraphs are dedicated to delineating the scope of data acquisition and preparatory measures employed to enrich the CENED dataset with readily and effortlessly accessible additional features. A model based solely on such features would be useful for any building, provided fundamental cadastral and geographic information is available.

3.3.1 Osservatorio OMI

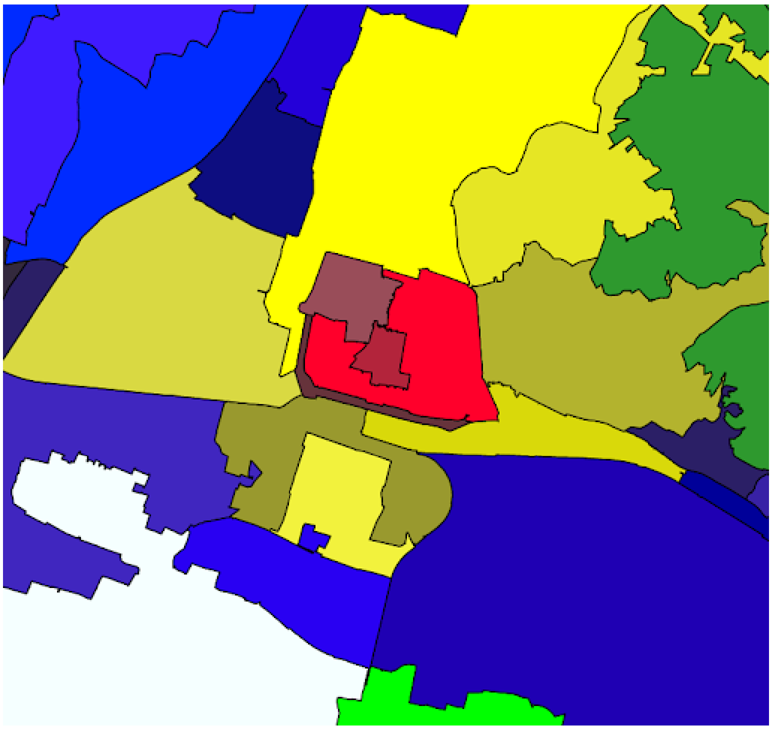
The Real Estate Market Observatory ("*Osservatorio del Mercato Immobiliare*", OMI) is an agency managed by *Agenzia delle Entrate* which collects and publishes several statistics on the real estate market on a 6-month basis, including maximum and minimum market prices and rents.

Data are provided in the so-called "OMI zones", defining distinct areas of the survey. Each zone is essentially a partition of a map (polygons) that can be regarded as being homogeneous as regards the type of urban landscape (available types are Central, Semi-central, Outer, Suburban, and Rural). This dataset provides the average market value for each cadastral entity at the OMI zone level.

building using non-renewable energy, which, compared to a theoretically similar unit with standard energy-saving features, defines the energy certificate class (from G to A4)

[24] In 2015, the government published new national guidelines for the EPC, leading to a discontinuity with those issued before.

Figure 1: OMI zones around the city center of Brescia in Lombardy.



Note: OMI zones around the city center of Brescia. Central zones are red, Semi-central zones are yellow, Suburban zones are blue, and Rural zones are green.

3.3.2 ANAIP Data

The heating degree-per-day ("*gradi-giorno*" or HDD) variable represents the difference between the average indoor (set to 20° C) and the outdoor temperature in the municipality, summed over each day of the year. This information, published in the D.P.R. 412/93, was extracted from an open table provided by the "*Associazione Nazionale Amministratori Immobiliari Professionisti*" (ANAIP), the Italian National Association of Building Managers. This variable is higher in areas with harsher climates and thus serves as a coarse measure of the expected energy requirements for heating.

3.3.3 Other data

We also gathered some data from the Italian National Statistical Office (ISTAT). ISTAT provides several statistical aggregates at the municipal level, including population, population trends, and per-capita income trends.

Supplementary data that have the potential to enhance the predictive power of an energy efficiency model are accessible for extensive geographical aggregates, specifically at the level of provinces:

- Natural gas consumption by province in 2020 (m^3, source: Italian Ministry for Ecological Transition - MITE).
- Mean household electricity consumption by province in 2020 (source: TERNA[25] Group's statistical annual report).
- Solar energy production by province in 2020 (source: TERNA Group's statistical annual report).

These data are used to improuve the accuracy of the model (see section 4.3).

4 Statistical Methodology applied

The aim of this project is to develop a machine learning model for predicting energy classes. Before delving into its technical details, a couple of considerations need to be accurately discussed.

Firstly, it's important to emphasize that an energy class prediction model is only useful when applied to buildings without certification, i.e., those not covered by the CENED dataset. For this reason, CENED data have only been used to extract the target variable, namely the energy class, while input variables have been selected from complementary, freely, and easily accessible data sources listed in Section 3. A model based solely on these features will be valuable for any building, as long as

[25] TERNA is the national operator managing high-voltage transmission grids all over Italy.

basic cadastral and geographic information is available. Additional variables have been created using data engineering techniques, as will be shown in Paragraph 4.2.

Furthermore, to reduce training time during the parameter optimization step, the model was tuned on a geographic subset of the data (Brescia province), as detailed in Section 4.1. This strategy is also useful for testing generalization capabilities and ascertaining if the model is relying too heavily on specific characteristics found in certain areas.'

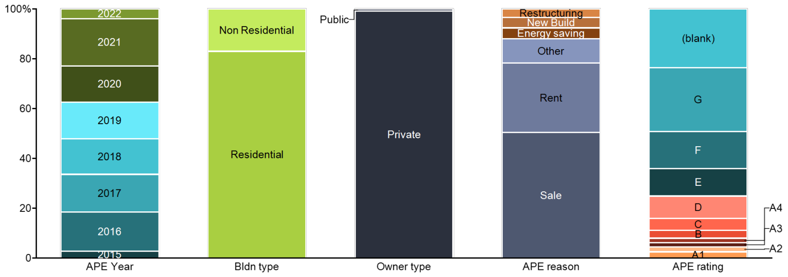
4.1 Training subset

Lombardy is the Italian region with the highest number of certificates available. The province of Brescia was chosen as the training subset due to the following considerations:

- it is the largest province in Lombardy (4,786 Km^2);
- it has 203 municipalities spanning a large range of population values, from Brescia ($\sim$ 200,000 inhabitants) to villages with less than 3,000 residents, totaling 1.2 million residents as of 2022;
- it has a wide variety of land types (mountains, valleys, hills, lakeside);
- being one of the most industrial areas in Italy, it should well represent all types of buildings (residential, hotels, offices, factories, warehouses, etc.), with a total of 1.8 million units;
- it has a significant amount of energy certificates available, totaling 152,000 certificates (out of the 1.21 million of our sample) at the time of this work.

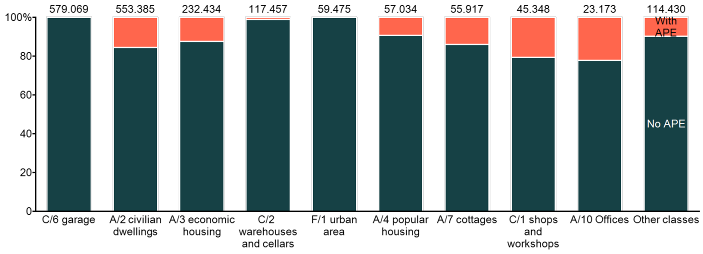
Figures 2, 3, and 4 show the distribution of the energy certificates collected in the province of Brescia and how they are distributed among cadastral categories and main municipalities:

Figure 2: EPCs in the Brescia province - overview.



Note: energy class is missing in 20% of the EPCs. These observations have been dropped. Source: CENED.

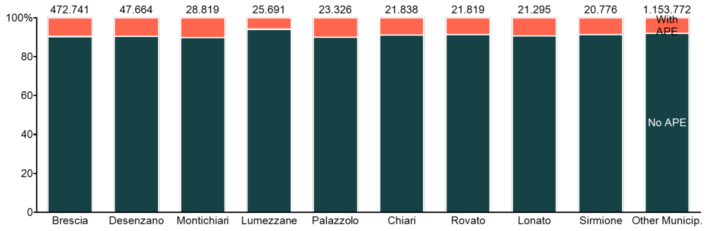
Figure 3: Buildings in the Brescia province - by building type



	C/6 garage	A/2 civilian dwellings	A/3 economic housing	C/2 warehouses and cellars	F/1 urban area	A/4 popular housing	A/7 cottages	C/1 shops and workshops	A/10 Offices	Other classes
% class on overall	32%	30%	13%	6%	3%	3%	3%	2%	1%	6%
% APE within class	0%	16%	12%	1%	0%	9%	14%	21%	22%	10%
% APE on total APE	0%	55%	18%	1%	0%	3%	5%	6%	3%	7%

Note: the total of all columns equals the total number of units per type, for a total of 1.8 million units. Sources: CENED and our elaborations on the Italian Cadastre.

Figure 4: Buildings in the Brescia province - by municipality.



	Brescia	Desenzano	Montichiari	Lumezzane	Palazzolo	Chiari	Rovato	Lonato	Sirmione	Other Municip.
% class on overall	26%	3%	2%	1%	1%	1%	1%	1%	1%	63%
% APE within class	10%	10%	10%	6%	10%	9%	9%	9%	9%	8%
% APE on total APE	29%	3%	2%	1%	2%	1%	1%	1%	1%	59%

Sources: CENED and our elaborations on the Italian Cadastre

4.2 Feature Engineering: year of construction

To exploit the maximum amount of information from the available data sources, feature engineering techniques have been applied to extract the year of construction for each building.

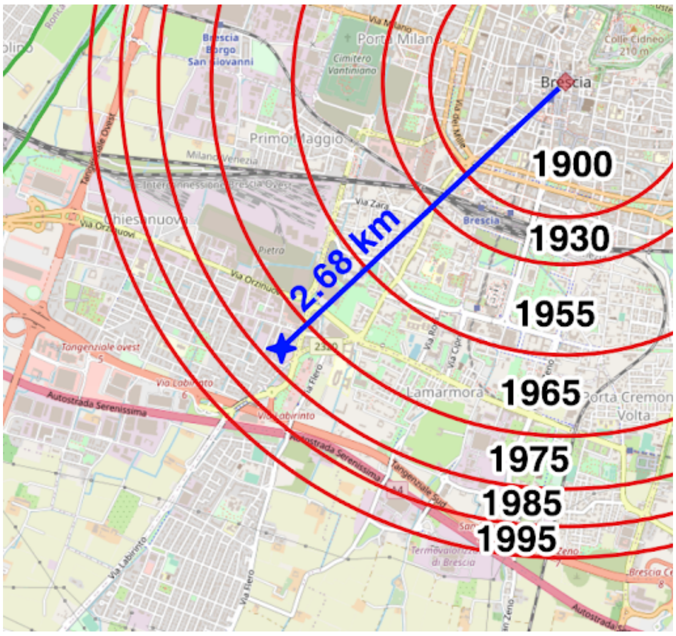
Indeed, the year of construction of a given building is likely a relevant feature in predicting energy classes.[26] Unfortunately, this variable isn't freely accessible and thus not included in the *Elenco Immobili* dataset. For this reason, a geometric model has been adopted to estimate the year of construction of buildings based on their distance from the old town, as detailed in Figure 5.

Urban areas typically expand radially from their center, assuming that most new buildings occupy previously vacant areas. In this approximation, urban expansion can be represented as a circle whose radius increases over time, with the center of the circle often corresponding to the old town. If the rate of radial expansion can be characterized using historical data, the year of construction can be inferred from the building's position relative to the circle's center.

ISTAT provides census data on residential buildings constructed in every municipality across various "eras", including counts in 1900, 1930, 1955, 1965, 1975, 1985, 1995, 2002, and 2008. In the absence of additional data, the conversion of residential building census data into urban surface areas has been based on a heuristic approach. This approach relates the density of residential buildings, a clear indicator of urbanization, to a higher fraction of urban areas. The density of residential buildings was approximated as the latest residential building count divided by the municipality's surface area, while the corresponding fractions of the urban area were determined with the assistance of domain experts and are listed

[26] The first law promoting energy efficiency in Italy was published in 1976 (law 373/76).

Figure 5: Representation of our model for estimating the year of construction of a building from its location



in Table 1. These fractions were heuristically adjusted to align with the ISTAT urbanization scale in the Lombardy region. They were then used to recalculate the residential building density in the urban area, which, in turn, was used to convert the residential building count for a given year into the urban surface area.

Table 1: Estimates for the urban area fraction against residential density

Residential Density (buildings / km^2)	Fraction of urban area
70	10%
100	25%
200	35%
400	50%
>400	70%

Finally, for each municipality, the urban epicenter was determined using street address geolocation via the Geoapify service, employing toponyms as search keywords. This method is preferred over calculating the centroid of GIS polygons, as it is not guaranteed to be close to the true center of the urban area. The "urban disc" is expected to be much smaller than the municipality boundaries, resulting in many buildings falling outside of it. For these peripheral buildings, the average year of construction of all residential buildings was utilized. The output of this processing step is a table that lists the urban area of each municipality for each "era", along with its epicenter and the average year of construction. This data is then utilized to predict the year of construction for each building based on its geolocation.

4.3 Model development and training

To choose and tune the optimal algorithm for the energy class prediction task, hyperparameter optimization was applied to several algorithms, namely Random Forest Classifier (RFC), Random Forest Regressor (RFR), and XGBoost (XGB). The Brescia subset was divided into training and test sets using a 75%-25% random split, while the parameter optimization procedure was evaluated using 5-fold cross-validation.[27] Table 2 lists all the variables used in this step, along with their source.
Besides usual classification metrics (accuracy, precision, recall, F1-score), two further quantities are used to better capture task-specific performance for the models, namely:

- **numerical deviation**: average of the absolute value of deviation from true values, expressed as the number of energy classes (e.g. true value = D, prediction = F, numerical deviation = 2)
- **1-class accuracy**, evaluated with a tolerance band of one class, i.e. accepting predictions that are off by one class.

In particular, 1-class accuracy is intended to give a more benign picture of the actual performance, as slightly off predictions can pragmatically be regarded as useful in evaluating a stock of buildings.

The performance criterion for the choice of the best model was **accuracy**. The latter selected a Random Forest Classifier (RFC) whose hyperparameters are listed in table 3. Model performance on the test set is reported in table 4.

The model trained so far experiences a 4 percentage points drop in 1-class accuracy and a 6 p.p. drop in accuracy when applied to the whole Lombardy dataset. This worsening denotes poor generalization capabilities of the model and can be attributed to different provincial features not explicitly included in the dataset, such as differences in building techniques, materials, survey strategies, and conventions.

Freezing hyperparameters to their optimal values (table 3), the RFC algorithm has been retrained on the full Lombardy dataset including the province-level features mentioned in section 3.3.3, resulting in overall improved metrics, as shown in table 5. Model explainability (as provided by variable importance) ranks surface area and market value on top for determining energy efficiency as shown in table 6. The year of construction, extracted with procedure described in section 4.2, also figures among the most important variables.

[27] N-fold Cross-Validation is a Machine-Learning technique that evaluates N models trained over a subset of the training set, which is different for each model. The average performance of the N models is the score used to optimize the set of hyper-parameters.

Table 2: List of input variables

Feature	Data source
Year of construction	own elaborations
Unit surface	Cadastre + Elenco_immobili
Floor	Cadastre + Elenco_immobili
Floors in building	Cadastre + Elenco_immobili
Number of units per floor	Cadastre + Elenco_immobili
Boundary unit	Cadastre + Elenco_immobili
Cadastral class (12 features)	Cadastre
Omi class (5 features)	OMI
Value per sqm	OMI
Province	ISTAT
Region	ISTAT
Heating degree days	ANAIP

The output class distribution displays a tendency to over-predict class G, as easily visualized in figure 6. An attempt was performed by repeating the whole training procedure on a balanced input set, however, it didn't yield encouraging results. As an alternative solution, the true distribution of energy classes obtained from SIAPE data was used to introduce a probabilistic correction to the model output. This correction is based on gain coefficients computed on the training set depending on province (p) and energy class (c). Taking the total number of predictions of class c in province p, $n_{c,p}^{Model}$, a posterior probability is computed as:

$$p_{c,p}^{Model} = n_{c,p}^{Model} / n_{p}^{Training} \tag{1}$$

being $n_p^{Training}$ the total number of observation belonging to province p in the training set. Similarly, an empirical distribution is inferred from SIAPE data $p_{c,p}^{SIAPE}$. Then a class- and province-dependent deviation of the model distribution from the empirical one is computed as:

$$dev_{c,p} = p_{c,p}^{SIAPE} - p_{c,p}^{Model} \tag{2}$$

Table 3: Optimal hyperparameters

RF Classifier hyper-parameters	
max_features	0.5
max_samples	0.5
n_estimators	200
max_depth	20
min_samples_split	5
min_samples_leaf	1

Table 4: Best model - performance metrics on the Brescia subset

RF Classifier metrics	
Numerical Deviation	1.43 ± 1.87
1-class accuracy	67.5%
Accuracy	40.0%
Macro-precision	48.0%
Macro-Recall	28.0%
Macro-F1 score	32.0%

Table 5: Performance - Lombardy

RF Classifier metrics	
Numerical Deviation	1.26 ± 1.65
1-class accuracy	70.0%
Accuracy	40.0%
Macro-precision	54.0%
Macro-Recall	30.0%
Macro-F1 score	35.0%

Table 6: Feature importance

Feature	Importance
Surface area	0.22
Market value	0.15
Sub-units per floor	0.14
Degrees per day	0.10
Year of construction	0.09
Floor	0,06

This allows us to define a probability gain coefficient for all predictions belonging to province p:

$$gain_{c,p} = 1 + \gamma \frac{dev_{c,p}}{\sum\limits_{c=A1,A2,...G} |dev_{c,p}|} \tag{3}$$

These coefficients are used to correct the class probability vector of each training observation, while the value of γ is optimized to maximize 1-class accuracy, which is achieved at γ = 1.

Table 7 shows that this procedure does not improve single-class metrics, which is reasonable since the correction results in a sub-optimal algorithm. However, significant improvements are achieved for numerical error and 1-class accuracy, indicating that predictions are on average closer to the true class, as shown in the confusion matrix in Figure 7. These corrections are only feasible for those provinces that are adequately represented in the training set so that reliable values for $p_{c,p}^{Model}$ can be calculated. Furthermore, the province must be present in the SIAPE survey, which isn't always the case e.g. in Campania or Sardinia. In the following, this correction will be adopted whenever data guarantee its validity.

To assess robustness, we conducted the same analysis using a subset (approxi-

Figure 6: Confusion matrix bar-plot for the RF Classifier with the Province model.

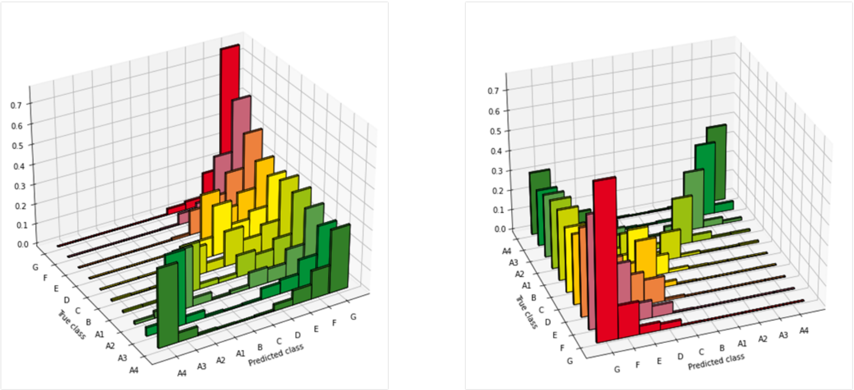
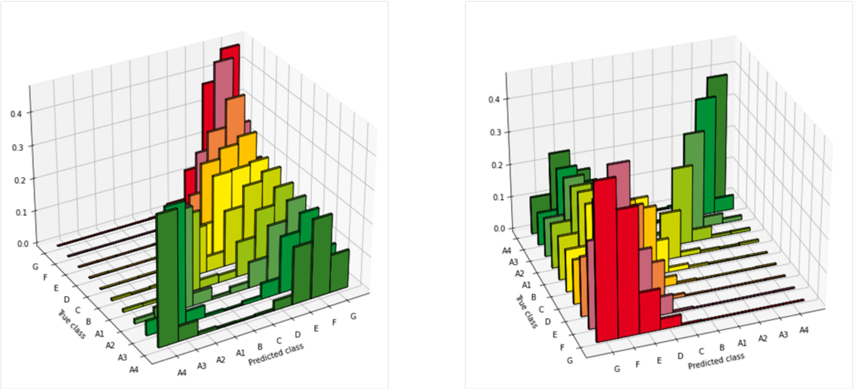


Figure 7: Confusion matrix bar-plot for the RF Classifier with probability gain.



mately 50%) of the EPC archive from Piedmont. The results, shown in Table 8, underline the significance of utilizing probability gains to enhance the performance of the RFC. Notably, when using the Piedmont archive, we observe a slight decrease in accuracy. This underscores the presence of regional heterogeneity and emphasizes the importance of accessing the whole microdata of the SIAPE for improved forecasting capabilities.

4.4 Classical reference - Ordered logit model

To demonstrate the additional value brought in by machine learning techniques, we conducted an estimation using a parametric approach known as a discrete choice model. Given the nature of the setting, we employed an ordered logit model (for a review of the model, refer to Wooldridge 2002) and utilized the dataset corresponding to the entire Lombardy region. This technique is particularly suitable for the task at hand as the energy classes adhere to the proportional odds assumption upon which the model is based. Furthermore, the model offers distinct advantages in terms of the interpretability of regression coefficients. Additionally, a parametric approach provides confidence intervals, variance estimates, and other metrics that

Table 7: Performance of RF Classifier with provincial features and probability gains

RF Classifier with Province metrics	
Numerical Deviation	1.19 ± 1.49
1-class accuracy	73.7%
Accuracy	37.0%
Macro-precision	54.0%
Macro-Recall	30.0%
Macro-F1 score	36.0%

Table 8: Robustness check: RFC trained on Lombardy vs. Piedmont

RF Classifier with probability gains		
	Lombardy	**Piedmont**
Numerical Deviation	1.19 ± 1.49	1.74 ± 1.9
1-class accuracy	73.7%	61.3%
Accuracy	37.0%	25.0%
Macro-precision	54.0%	14.0%
Macro-Recall	30.0%	13.0%
Macro-F1 score	36.0%	12.0%

are not available in a non-parametric approach.

The ordered logit model achieves an accuracy of 32.7% and a 1-class accuracy of 62.9%. This represents a performance loss of approximately 7 percentage points compared to Table 5 and between 4 and 10 p.p. once introducing data-driven probability gains (see table 7).

The superior performance of machine learning models in this setting can be attributed to multiple factors. First, Random Forest is capable of capturing nonlinear effects that may be present in such a complex task. Moreover, the dataset exhibits a significant class imbalance, which is known to have a detrimental impact on linear models, affecting both performance and parameter estimates. Random Forest, based on ensemble learning and sampling (bagging), has the ability to downsample the majority class and construct trees on a more balanced dataset.

5 Results

Assessment of the energy class prediction model obtained so far is not straightforward, mainly due to the lack of an extensive, energy-efficiency labeled dataset. In the next sections the model will be tested against:

- SIAPE dataset. Due to the lack of microdata, we compare our results with the regional class distributions provided on the website of SIAPE;
- *bank X* dataset.[28] We compare our results to the EPCs collected by *bank X* from its clients.

5.1 Comparison to the SIAPE dataset

The resulting model was applied to a dataset of 1.49 million buildings chosen randomly[29] as per Figure 8. Data are primarily located in the regions of Lombardy,

[28] A portfolio of loans from an Italian bank was obtained and used confidentially by Datasinc.

[29] The dataset size was adjusted to be as large as possible while containing the computational and temporal cost of the analysis. Trentino Alto-Adige was excluded.

Veneto, Emilia-Romagna, and Lazio. As previously mentioned, the SIAPE dataset is incomplete and does not provide, yet, EPC information for 2 out of 21 regions (Campania and Sardinia). For this reason, buildings located in those provinces could not benefit from the probability gain method described in Equation 3.

The lack of microdata in this dataset limits validation procedures to a qualitative cross-check of the distributions of model predictions w.r.t. those found in the SIAPE dataset, focusing on residential properties (which accounts for almost 90% of the EPCs in the SIAPE). Figures 9 and 10 report the EPCs distribution values by region, while figure 11 gives an overview of the deviations between the models and the SIAPE dataset at the national level. The model shows a tendency to overpredict one of the least energy efficient classes (F) while underpredicting the other ones. Deviations are shown in absolute magnitude for better visualization. It's worth noting that according to our model houses belonging to the bottom of the energy distribution (classes G and F) would be 82% of the total w.r.t. 56% which could be inferred from SIAPE, indicating that there are 9 million more houses that need retrofitting. This is significant in terms of policy implications, which we will discuss later.

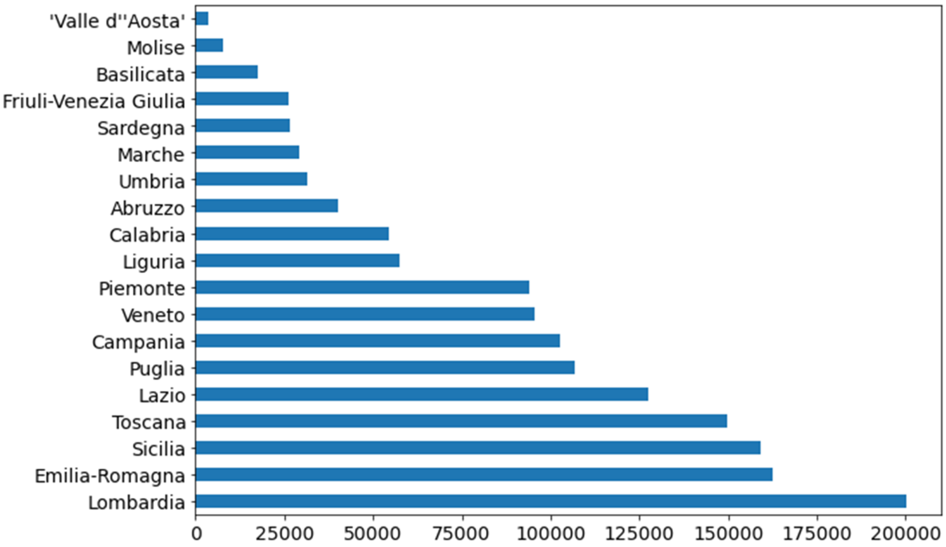
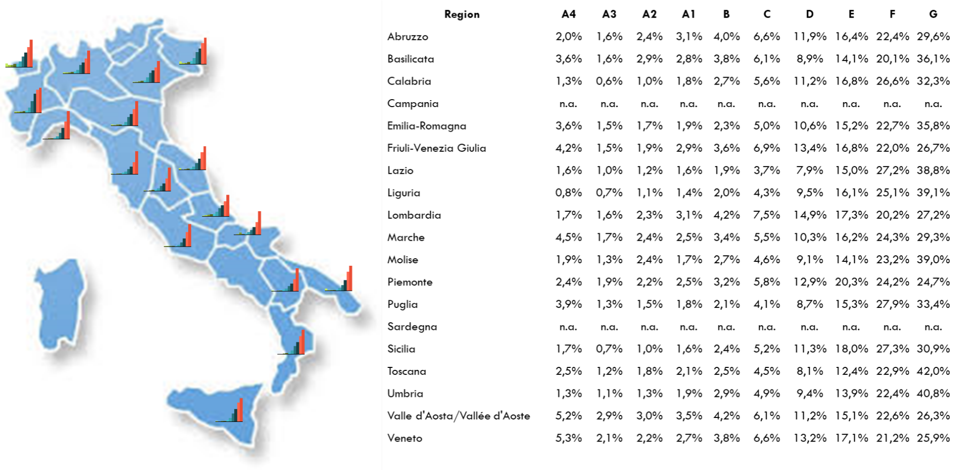
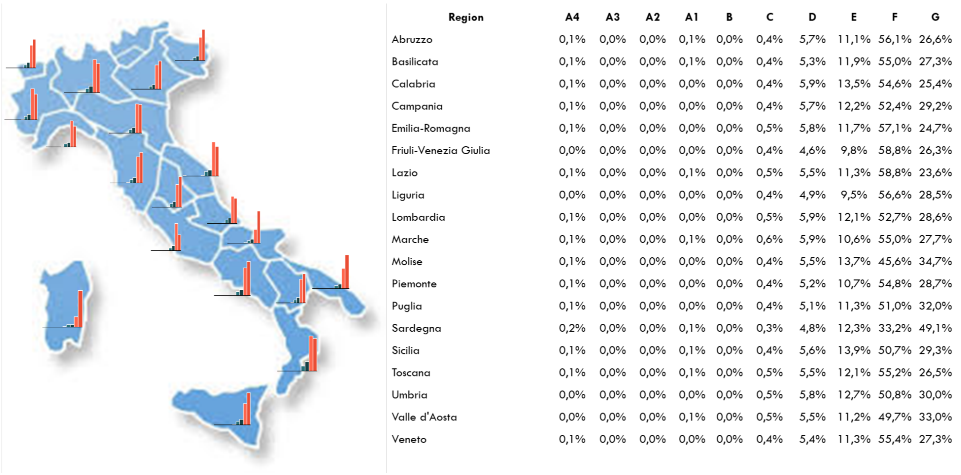


Figure 8: Building distributions by region for the Italy dataset



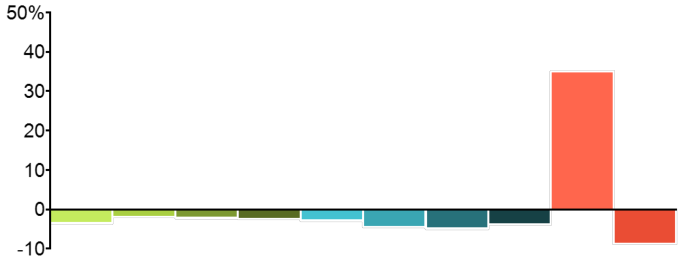
Region	A4	A3	A2	A1	B	C	D	E	F	G
Abruzzo	2,0%	1,6%	2,4%	3,1%	4,0%	6,6%	11,9%	16,4%	22,4%	29,6%
Basilicata	3,6%	1,6%	2,9%	2,8%	3,8%	6,1%	8,9%	14,1%	20,1%	36,1%
Calabria	1,3%	0,6%	1,0%	1,8%	2,7%	5,6%	11,2%	16,8%	26,6%	32,3%
Campania	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Emilia-Romagna	3,6%	1,5%	1,7%	1,9%	2,3%	5,0%	10,6%	15,2%	22,7%	35,8%
Friuli-Venezia Giulia	4,2%	1,5%	1,9%	2,9%	3,6%	6,9%	13,4%	16,8%	22,0%	26,7%
Lazio	1,6%	1,0%	1,2%	1,6%	1,9%	3,7%	7,9%	15,0%	27,2%	38,8%
Liguria	0,8%	0,7%	1,1%	1,4%	2,0%	4,3%	9,5%	16,1%	25,1%	39,1%
Lombardia	1,7%	1,6%	2,3%	3,1%	4,2%	7,5%	14,9%	17,3%	20,2%	27,2%
Marche	4,5%	1,7%	2,4%	2,5%	3,4%	5,5%	10,3%	16,2%	24,3%	29,3%
Molise	1,9%	1,3%	2,4%	1,7%	2,7%	4,6%	9,1%	14,1%	23,2%	39,0%
Piemonte	2,4%	1,9%	2,2%	2,5%	3,2%	5,8%	12,9%	20,3%	24,2%	24,7%
Puglia	3,9%	1,3%	1,5%	1,8%	2,1%	4,1%	8,7%	15,3%	27,9%	33,4%
Sardegna	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Sicilia	1,7%	0,7%	1,0%	1,6%	2,4%	5,2%	11,3%	18,0%	27,3%	30,9%
Toscana	2,5%	1,2%	1,8%	2,1%	2,5%	4,5%	8,1%	12,4%	22,9%	42,0%
Umbria	1,3%	1,1%	1,3%	1,9%	2,9%	4,9%	9,4%	13,9%	22,4%	40,8%
Valle d'Aosta/Vallée d'Aoste	5,2%	2,9%	3,0%	3,5%	4,2%	6,1%	11,2%	15,1%	22,6%	26,3%
Veneto	5,3%	2,1%	2,2%	2,7%	3,8%	6,6%	13,2%	17,1%	21,2%	25,9%

Figure 9: SIAPE aggregated class distributions by region



Region	A4	A3	A2	A1	B	C	D	E	F	G
Abruzzo	0,1%	0,0%	0,0%	0,1%	0,0%	0,4%	5,7%	11,1%	56,1%	26,6%
Basilicata	0,1%	0,0%	0,0%	0,1%	0,0%	0,4%	5,3%	11,9%	55,0%	27,3%
Calabria	0,1%	0,0%	0,0%	0,0%	0,0%	0,4%	5,9%	13,5%	54,6%	25,4%
Campania	0,1%	0,0%	0,0%	0,0%	0,0%	0,4%	5,7%	12,2%	52,4%	29,2%
Emilia-Romagna	0,1%	0,0%	0,0%	0,0%	0,0%	0,5%	5,8%	11,7%	57,1%	24,7%
Friuli-Venezia Giulia	0,0%	0,0%	0,0%	0,0%	0,0%	0,4%	4,6%	9,8%	58,8%	26,3%
Lazio	0,1%	0,0%	0,0%	0,1%	0,0%	0,5%	5,5%	11,3%	58,8%	23,6%
Liguria	0,0%	0,0%	0,0%	0,0%	0,0%	0,4%	4,9%	9,5%	56,6%	28,5%
Lombardia	0,1%	0,0%	0,0%	0,0%	0,0%	0,5%	5,9%	12,1%	52,7%	28,6%
Marche	0,1%	0,0%	0,0%	0,1%	0,0%	0,6%	5,9%	10,6%	55,0%	27,7%
Molise	0,1%	0,0%	0,0%	0,0%	0,0%	0,4%	5,5%	13,7%	45,6%	34,7%
Piemonte	0,1%	0,0%	0,0%	0,0%	0,0%	0,4%	5,2%	10,7%	54,8%	28,7%
Puglia	0,1%	0,0%	0,0%	0,0%	0,0%	0,4%	5,1%	11,3%	51,0%	32,0%
Sardegna	0,2%	0,0%	0,0%	0,1%	0,0%	0,3%	4,8%	12,3%	33,2%	49,1%
Sicilia	0,1%	0,0%	0,0%	0,1%	0,0%	0,4%	5,6%	13,9%	50,7%	29,3%
Toscana	0,1%	0,0%	0,0%	0,1%	0,0%	0,5%	5,5%	12,1%	55,2%	26,5%
Umbria	0,0%	0,0%	0,0%	0,0%	0,0%	0,5%	5,8%	12,7%	50,8%	30,0%
Valle d'Aosta	0,0%	0,0%	0,0%	0,1%	0,0%	0,5%	5,5%	11,2%	49,7%	33,0%
Veneto	0,1%	0,0%	0,0%	0,0%	0,0%	0,4%	5,4%	11,3%	55,4%	27,3%

Figure 10: Aggregated model predictions by region



	A4	A3	A2	A1	B	C	D	E	F	G
SIAPE - Residential	4%	2%	2%	3%	3%	5%	11%	16%	24%	32%
Best fitting model Estimate - Residential	0%	0%	0%	0%	0%	0%	6%	12%	59%	23%

Figure 11: Deviation in predictions vs. SIAPE data for national residential properties

5.2 Comparison to *bank X* data

In order to test the quality of the model, *bank X* provided a dataset containing part of its real estate collaterals located in Lombardy.[30] This dataset includes the details of each property (Municipality, “Foglio”, “Mappale” and “Subalterno” when available) and the energy rating collected during the loan agreement process (if available). The portfolio contains approximately 169 thousand units, including those not required to post an EPC (e.g. garages/boxes or land properties). Considering that the collection of energy certificates in real estate transactions was introduced in 2010, most observations do not include information about the energy class. As a result of these limitations, *bank X* dataset provides an energy class for 33.6 thousand units (20% of observations), but only 15.5 thousand have an energy rating. Table 9 shows the performance and accuracy of the model on the overall portfolio. To correctly interpret these results, it is worth to remark the following:

- the energy rating scale has undergone several changes through time. In particular, before 2015 it encompassed labels “A” or “A+”, while starting from 2016 it has been divided into four classes, from "A4" to "A1". In the current evaluation all “A” and “A+” have been considered correct if mapped in the new A1-A4 classes - see the first row in table 9;
- for completeness the model has also been run on units for which *bank X* has no available energy class rating - see row "No Data" (60,686 EPCs estimated but with no real EPCs to compare with);
- for 92 thousand units in the portfolio the model does not provide an energy rating: 76 thousand are units that don’t require an energy rating (e.g. garages, attics); 3 thousand are units in the cadastral classes for which the model has not been trained (e.g. cadastral type F); 14 thousand are not provided due to issues in the raw data.

[30] Data were retrived and used by Datasinc.

Table 9: Confusion matrix between truth and our model's predictions for the *bank X* dataset

Energy class provided by *bank X*	Energy class estimate with best fitting model											
	A4	**A3**	**A2**	**A1**	**B**	**C**	**D**	**E**	**F**	**G**	**No Estim.**	**Total**
A	2	0	0	3	0	5	16	106	140	40	826	**1,138**
A+	1	0	0	0	0	0	3	24	14	12	165	**219**
A4	4	1	0	0	0	0	4	32	75	19	396	**531**
A3	0	4	0	0	0	0	3	37	59	13	367	**483**
A2	0	0	2	0	0	1	13	40	54	16	346	**472**
A1	0	0	3	1	0	0	7	46	77	24	376	**534**
B	11	0	3	0	12	0	84	211	449	144	1,519	**2,433**
C	0	0	1	0	0	3	54	177	343	136	1,029	**1,743**
D	2	0	0	0	0	0	75	293	722	288	1,725	**3,105**
E	0	2	0	2	0	3	81	542	1,313	533	2,816	**5,292**
F	2	0	0	0	0	1	74	487	2,006	878	3,525	**6,973**
G	0	0	1	0	1	0	109	631	2,784	2,137	5,087	**10,750**
No Data	96	54	36	48	66	275	3,462	10,452	30,066	16,131	74,457	**135,143**
Total	**118**	**61**	**46**	**54**	**79**	**288**	**3,985**	**13,078**	**38,102**	**20,371**	**92,634**	**168,816**

When summarizing the results achieved on this dataset (Table 10), and focusing on the units for which microdata are available (15.5 thousand units), it turns out that accuracy is 31% while 69% of the predictions match the corresponding true value within one class (1-class accuracy). The majority of the "errors" are due to underestimation of the energy class by the model, in coherence with what was already observed while assessing the model w.r.t. SIAPE data.

Table 10: Class-match deviations

	Absolute #	**% on total**
Perfect match	4,792	31%
1 class	5,894	38%
2 classes	2,227	14%
Not matched	2,583	17%
Total predictions	15,496	100%

6 Conclusions

The decarbonization of the building sector is crucial for achieving the EU's climate targets. Detailed information on buildings' energy performance is essential to guide investment decisions and assist the financial sector in reducing its carbon footprint and complying with regulatory requirements. Several European countries, including France[31] and Spain[32], maintain centralized, open-data archives of

[31] The French Environment and Energy Management Agency (ADEME) publishes its EPC dataset as Open Data, likely in accordance with the Open Data general rules and Article L. 126-32 of the French Construction Code.

[32] In Spain, EPC regional datasets are included in the Open Data National Platform as Spanish law restricts access to the Spanish Cadastral and Property Registries.

their EPCs. In Italy, however, each region is responsible for collecting EPC data, which are mandated by law only for buildings sold, rented, or refurbished since 2010. These regional energy performance cadastres are often inaccessible, except for Lombardy, Piedmont, and Trentino.

ENEA is currently gathering data from various regions to populate a national dataset, SIAPE, but the microdata remains unavailable.

In this paper, we present a predictive model for determining the energy class of all building portions across Italy. Our model is based on a Random Forest Classifier trained using readily available data from Lombardy, which is Italy's most populous and affluent region. The model identifies surface area, market value, year of construction, heating degree days, and the number of sub-units per floor as the most significant variables for predicting energy classes. Notably, all these variables are easily retrievable by banks.

Our model achieves a 37% accuracy rate in correctly predicting the energy class. When allowing for a one-class margin of error (where a "real class C" might be mistakenly labelled as either a "B" or a "D"), the accuracy increases to 74%. Furthermore, when applying our model to data from the only other available regional open data source, Piedmont, we observe a slight decrease in accuracy. This highlights the presence of regional heterogeneity in the regulations governing EPCs, as also noted by Loberto et al. (2023), who analyzed a large dataset of real estate listings on the Immobiliare.it platform, Italy's largest online portal for real estate services.

Additionally, we compare the results with the regional distribution of EPCs from SIAPE and with a proprietary database from an Italian bank.

In comparison to SIAPE, particularly within the residential sector, our model demonstrates an overestimation of the proportion of least energy-efficient homes (classes G and F) by nearly 26 percentage points, equating to approximately 9 million additional houses requiring energy retrofitting in the near future. This discrepancy serves as a significant alarm for policymakers aiming to craft effective energy efficiency policies. While our estimates may not precisely reflect the exact number of least energy-efficient homes, it underscores the urgency for further investigation given the policy implications of this issue. Moreover, it's important to note the simplicity of our model, which relies on only a few readily available variables, in contrast to the numerous parameters required to generate EPCs.

Also, although modest, the accuracy of our model represents a notable improvement compared to a traditional parametric approach, such as an ordered logit, which achieves lower accuracy.

To obtain a more accurate assessment of energy efficiency in buildings, microdata from SIAPE should be made publicly accessible, and the coverage ratio should be increased, meaning more buildings should undergo energy audits.

The current performance of our model can be viewed as a conservative estimate, as evidenced by other research (Mayer et al., 2023) demonstrating the positive impact of satellite and street-view imagery on similar models. Incorporating freely available satellite imagery, such as that provided by the Copernicus Programme, represents an intriguing avenue for future development and is likely to result in improved performance.

References

Banca d'Italia. Survey on Household Income and Wealth - 2020. Technical report, Banca d'Italia, 2022. URL https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/bil-fam2020/Fascicolo_IBF_2020_ENG.pdf?language_id=1.

BCBS. Climate-related risk drivers and their transmission channel. Technical report, Basel Committee on Banking Supervision, 2021.

Enrico Bernardini, Ivan Faiella, Luciano Lavecchia, Alessandro Mistretta, and Filippo Natoli. Central banks, climate risks and sustainable finance. Questioni di Economia e Finanza (Occasional Papers) 608, Bank of Italy, Economic Research and International Relations Area, March 2021. URL https://ideas.repec.org/p/bdi/opques/qef_608_21.html.

European Commission. Regulation (EU) 2020/852 on the establishement of a framework to facilitate sustainable investment. Technical report, European Commission, 2020.

European Commission. Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) no 575/2013 as regards requirements for credit risk, credit valuation adjustment risk, operational risk, market risk and the output floor. Technical report, European Commission, 2021.

EBA. Guidelines on loan origination and monitoring. Technical report, European Banking Authority, 2020.

EBA. Report on management and supervision of ESG risks for credit institutions and investment firms. Technical report, European Banking Authority, 2021.

EBA. Final draft implementing technical standards on prudential disclosures on ESG risks in accordance with Article 449a CRR. Technical report, European Banking Authority, 2022.

EBA. Report on the role of environmental and social risks in the prudential framework. Technical report, European Banking Authority, 2023.

ECB. Guide on climate-related and environmental risks. Technical report, European Central Bank, 2020.

Emad Elbeltagi and Hossam Wefki. Predicting energy consumption for residential buildings using ANN through parametric modeling. *Energy Reports*, 7: 2534–2545, 2021. ISSN 2352-4847. doi: https://doi.org/10.1016/j.egyr.2021.04.053. URL https://www.sciencedirect.com/science/article/pii/S2352484721002705.

European Commission. Action plan on sustainable finance. Technical report, European Commission, 2018.

European Commission. European green deal. Technical report, European Commission, 2019.

European Commission and Directorate-General for Energy. *The quantitative relationship between energy efficiency improvements and lower probability of default of associated loans and increased value of the underlying assets: final report on risk assessment*. Publications Office of the European Union, 2022. doi: doi/10.2833/532126.

Benjamin Guin and Perttu Korhonen. Does energy efficiency predict mortgage performance? Bank of England working papers 852, Bank of England, January 2020. URL https://ideas.repec.org/p/boe/boeewp/0852.html.

Xinyi Li and Runming Yao. Modelling heating and cooling energy demand for building stock using a hybrid approach. *Energy and Buildings*, 235:110740, 2021. ISSN 0378-7788. URL https://www.sciencedirect.com/science/article/pii/S0378778821000244.

Yaolin Lin, Jingye Liu, Kamiel Gabriel, Wei Yang, and Chun-Qing Li. Data-Driven Based Prediction of the Energy Consumption of Residential Buildings in Oshawa. *Buildings*, 12(11), 2022. ISSN 2075-5309. doi: 10.3390/buildings12112039. URL https://www.mdpi.com/2075-5309/12/11/2039.

Michele Loberto, Alessandro Mistretta, and Matteo Spuri. The capitalization of energy labels into house prices. Evidence from Italy. Questioni di Economia e Finanza (Occasional Papers) 818, Bank of Italy, Economic Research and International Relations Area, 2023.

Tommaso Loizzo and Federico Schimperna. ESG disclosure: regulatory framework and challenges for Italian banks. Questioni di Economia e Finanza (Occasional Papers) 744, Bank of Italy, Dec 2022. URL https://ideas.repec.org/p/bdi/opques/qef_608_21.html.

Kevin Mayer, Lukas Haas, Tianyuan Huang, Juan Bernabé-Moreno, Ram Rajagopal, and Martin Fischer. Estimating building energy efficiency from street view imagery, aerial imagery, and land surface temperature data. *Applied Energy*, 333, 2023.

NGFS. Guide for Supervisor. Technical report, Network for Greening the Financial System, 2020.

P. Reusens, F. Vastmans, and S. Damen. The impact of changes in dwelling characteristics and housing preferences on Belgian house prices. *Economic Review*, pages 1–40, April 2022. URL https://ideas.repec.org/a/nbb/ecrart/y2022mapril.html.

RICS. RICS Sustainability Report. Technical report, Royal Institution of Chartered Surveyors, 2022.

M. Sahu, Bishwajit Bhattacharjee, and Subhash C. Kaushik. Thermal design of air-conditioned building for tropical climate using admittance method and genetic algorithm. *Energy and Buildings*, 53:1–6, 2012. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2012.06.003. URL https://www.sciencedirect.com/science/article/pii/S0378778812002903.

Maoran Sun, Changyu Han, Quan Nie, Jingying Xu, Fan Zhang, and Qunshan Zhao. Understanding building energy efficiency with administrative and emerging urban big data by deep learning in Glasgow. *Energy and Buildings*, 273, 2022. doi: https://doi.org/10.1016/j.enbuild.2022.112331.

Sonia Williams and Christopher Bonham. Using machine learning to predict energy efficiency. Technical report, ONS, 2020. URL https://datasciencecampus.ons.gov.uk/projects/using-machine-learning-to-predict-energy-efficiency/.

Jeffrey M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2002.