THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Method development to improve estimates of embodied carbon emissions in the Swedish built environment

Qiyu Liu

Department of Space, Earth and Environment

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2025

Method development to improve estimates of embodied emissions in the Swedish built environment

Qiyu Liu

© Qiyu Liu, 2025.

Department of Space, Earth and Environment
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46(0)31-772 1000

Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Method development to improve estimates of embodied emissions in the Swedish built environment

QIYU LIU
Department of Space, Earth and Environment
Chalmers University of Technology

Abstract

The building and construction sector accounts for approximately 20 percent of Sweden's total greenhouse gas (GHG) emissions. To achieve Sweden's climate goal of reaching net-zero greenhouse gas emissions by 2045, urgent action is thus needed to decarbonize the building and construction sector. Emissions from the building sector can be classified into two sources: 1) operational emissions related to energy consumption, and 2) embodied emissions from the manufacturing and processing of building materials. The share of embodied emissions as a percentage of the total emissions is expected to increase as more efforts are put into energy efficiency measures and decarbonization of the energy supply. As a result, more attention is needed to reduce embodied emissions from the building and construction sector.

However, embodied emissions in the built environment remain relatively understudied, especially at the national level. Embodied emissions from the built environment are most commonly studied using material stock and flow analysis. Material stock and flow analysis can be classified into top-down and bottom-up, and this study focuses on the latter. Top-down analysis typically relies on aggregated national or sectoral data, such as economic or material flow accounts, while bottom-up analysis is based on detailed data at the building or component level. Existing bottom-up studies primarily focus on small geographical scales, such as neighborhood and city levels, which limits the applicability of their findings national level policy making. The primary challenge in conducting national-level estimations of embodied emissions lies in the limited availability of inventory data. Moreover, maintenance and renovation activities are frequently overlooked in models of material flows within the built environment.

This thesis addresses this gap by developing a framework to estimate embodied emissions from the built environment at the national level while having limited available data. The challenge of lack of available data is tackled by using machine learning (ML) models. Paper I estimates the material stock, flow, and embodied carbon from Swedish roads while predicting missing road widths data using a ML regression model. Paper II expands the scope to residential buildings by predicting construction years of buildings using a classification ML model and usable floor space of buildings using a regression ML model. Lastly, Paper III utilizes the building inventory dataset generated in Paper II to develop a material stock and flow model that introduces a new layered based approach to model renovation of buildings.

The findings presented in the appended work show that ML models can be used to predict physical attributes of roads and buildings to a high level of accuracy. The contribution of this work is showing that urban form features that can be generated using solely the geometry of roads and buildings can reliably achieve high level of prediction accuracy. Thereby increasing the applicability of the approach. The results also indicate that road maintenance and building renovation account for the largest share of embodied emissions. As a consequence, additional policy measures are needed to limit the emissions from maintenance and renovation activities.

Keywords: *Material stock, Material flow analysis, Machine learning, Embodied emissions*

List of publications

The thesis is based on the following appended papers, which are referred to in the text by their assigned Roman numerals:

I. Qiyu L, Johan R, Filip J. (2024) Development of a machine learning model to improve estimates of material stock and embodied emissions of roads. *Cleaner Environmental Systems*. https://doi.org/10.1016/j.cesys.2024.100211

II. Qiyu L, Maud L, Johan R, Filip J. (2025) Imputing missing data in building inventories: urban morphology indicators reliably predict age and floor are of buildings. *Smart and Sustainable Built Environment, Under review.*

III. Qiyu L, Maud L, Johan R, Filip J. (2025) Dynamic modeling of material flow and embodied emissions from building renovation – a layered material flow analysis approach. *Manuscript*.

Qiyu Liu is the principal author of all the appended papers. Professor Filip Johnsson contributed with discussion, methodology development, and editing of **Papers I-III**. Dr. Johan Rootzén contributed with discussion, methodology development, and editing of **Papers I-III**. Assistant professor Maud Lanau contributed with discussion, methodology development, and editing of **Papers II-III**.

Acknowledgements

This PhD journey would not have been possible without the support and encouragement of many people. To begin with, I would like to express my deepest gratitude to my supervision team. Filip, thank you for all the time and effort you have put in to discussing and reading my work. Johan, thank you for the wonderful ideas, 'sidetracks', and your willingness and patience to explore new methods with me. Maud, thank you for agreeing to becoming my co-supervisor. I have learned so much from you in a short period of time and this work would not have been possible without your support. I would also like to thank Ida, who is not officially part of the supervision team, but has contributed so much to my work.

This acknowledgement would not be complete without thanking the colleagues and friends at the Division of Energy Technology. I would like to start off by thanking our administrators, Marie, Anna, and Katarina for making sure everything runs smoothly. To the energy systems research group, thank you for all the valuable feedback and encouragement. In addition, I would like to thank all the incredible friends I have made during my time working in the division. I had no idea what to expect when I first stepped off the plane in Stockholm during COVID in 2021, but everything has turned out so much better than I could have ever imagined because of all of you. Anh, Anna, Farha, Georgia, Henrik, and I am sure I would miss some names, but thank you all so much for being on this journey with me. A special shoutout goes to Sina and Tharun, for making me partially a process engineer and to being there for me through all the highs and lows. I would also like to thank my colleagues at the Sustainable Built Environment Research Area at ACE.

Last but not least, I would like to thank my family for all of their support throughout all the years.

Qiyu Liu

Göteborg, May 2025

Table of Contents

1 Introduction

The building and construction sector is responsible for 34% of global energy demand and 37% of energy and process related greenhouse gas (GHG) emissions in 2022 [1]. At the same time, the global demand for housing and infrastructure is increasing at a rapid rate to meet the demand of a growing population. The Organization for Economic Co-operation and Development (OECD) estimated that the global construction sector will more than double by 2060, with its materials use reaching close to 84 gigatons (Gt) [2]. Therefore, decarbonizing the building and construction sector is essential for achieving the Paris Agreement targets and reaching net zero emissions [3]. In existing literature, a large amount of effort has been put into measures to reduce operational emissions from buildings from energy use [4]. However, as energy efficiency improves in buildings and energy supplies become more decarbonized, embodied GHG emissions from the manufacturing and processing of building materials are expected to become the largest source of emissions related to buildings [5]. In addition, road and rail infrastructures are responsible for approximately 0.5% - 1.9% of emissions globally in 2021 [6]. Together this highlights the need for increased focus on how to reduce embodied emissions from the entire built environment.

Embodied carbon emissions are directly related to the amount and type of materials used in the built environment, so understanding and estimating material flows and stocks is important for improving estimates of embodied carbon emissions [5]. Material stock models are essential to estimate where, and how much materials are accumulated, and are the basis to calculate material flows [7]. High spatial resolution material stock models can help policymakers make more targeted policy decisions, as there is high spatial heterogeneity in the built environment [8]. Material flow analysis results are used to provide insight into topics such as circularity and embodied emissions, and thus higher spatial resolution material stock and flow analysis (MSFA) models are needed to translate research into real world actions [7].

However, high-resolution, dynamic material flow analysis of the built environment are often constrained by the lack of data [7]. For buildings (the most studied stock), lack of data includes issues of missing data in the building inventory. Building inventory contains data such as dimensional information (e.g., floor space, volume) and archetype descriptors (e.g., construction year, structure type, building use) that are essential for calculating material stock and flows. Floor space data is used as the basis for calculating building stock and flows [9]. However, floor space data are often unavailable, and the available datasets are incomplete [10].

The lack of data also pertains to other parts of the built environment, with transport infrastructures being relatively understudied compared to buildings [11]. Transport infrastructures are an essential part of the built environment and the demand for new housing dictates the need for more transport infrastructure, and therefore it is crucial to consider the embodied emissions from transport infrastructures with regards to decarbonization of the built environment [12].

Furthermore, previous studies have largely focused on the dynamic material flows from new construction activities from stock expansion and replenishment (see **Paper III** for a review). The existing literature neglects renovation material flows to maintain the function and/or aesthetics of the stock. The reason that renovation activities have not been studied as much is similarly due to the lack of data. For example, the missing data are the data on the dimensions of the stock, year of construction, and detailed material intensity (MI) to understand which part of the stock needs to be renovated.

1.1 Aim and scope

The overall aim of this thesis is to improve estimations of material flows and embodied emissions from the Swedish built environment with a goal of decarbonization, with a specific focus on addressing issues of data availability and on increasing the level of modeling granularity. The geographical boundary of the study is Sweden, and the part of the built environment studied are residential buildings and roads.

The thesis addresses the following questions:

- Can machine learning methods be used to impute missing inventory data at the national level using limited input data? (RQ1)
- How to leverage higher-resolution stock models to improve estimations of maintenance/renovation activities and the associated material flows? (RQ2)
- What are the potentials to reduce embodied emissions from both new construction and maintenance/renovation activities? (RQ3)

1.2 Outline of the thesis

The thesis consists of a summary and three appended papers. The summary synthesizes the key findings from the papers and contextualizes the results. The summary begins with an introductory chapter, followed by Chapter 2 that provides background on the research topic.

Chapter 3 describes the methodology used, Chapter 4 presents some selected results from the works, and Chapter 5 provides discussion around the work. Lastly, Chapter 6 discusses the conclusions and future work from this work.

The first research question is addressed in **Papers I** and **II,** which both use machine learning methods to impute missing data in inventory datasets. In **Paper I,** a regression model is used to impute missing road width data. In **Paper II,** a regression model is used to impute missing residential building usable floor space data, and a classification model is used to estimate missing building construction years.

The second research question is addressed in **Papers I** and **III,** which both develop high spatial resolution material stock and flow models. In **Paper I,** road material stock is developed based on GIS dataset and the renovation of the top layer of roads is modeled using dynamic MFA. Similarly, **Paper III** models the material stock of residential buildings with detailed material intensity based on building shearing layers, and a method to model renovations of different shearing layers using dynamic MFA is developed.

The third research question is addressed in **Papers I** and **III,** which both estimate the embodied carbon from expansion and renovation activities of the material stocks. The inflow of material is multiplied by supply-chain scenario-based embodied emission factors from Karlsson et al. with a focus on Sweden [13]. The emission factors are developed to estimate the potential pathways for Sweden' construction industry to reach net zero emissions.

2 Background

This section provides some background information on the topics and methods that have been central to the work: *Material stock modeling of the built environment* (Section 2.1), *Machine learning to impute missing inventory data* (Section 2.2), *Dynamic material flow analysis to model maintenance/renovation* (Section 2.3), and *Sweden as a case study* (Section 2.4).

2.1 Material stock modeling of the built environment

Material stocks in buildings and infrastructures are a major source of lock-in that leads to path dependencies, and the stock determines the demand for raw materials and the outflows of waste and potentially recoverable resources. Therefore, the modeling of material stock is crucial to understand these dynamics and to better manage the built environment [14]. Higher spatial resolution of material stock and flow studies are needed to gain a deeper understanding of how to reduce embodied emissions, but the lack of quality data is a key limitation for conducting such studies [7]. This section underlines the importance of modeling the material stocks of roads and buildings and presents research and data gaps in existing literature.

2.1.1 Data availability challenges

Infrastructure construction, dominated by road construction, accounts for a significant share of the carbon footprint of the global construction sector. In 2013, Müller et al. [15] estimated the carbon footprint of the existing global infrastructure stock in 2008 as 122 (− 20/+ 15) $GtCO_2$. More recently, Rousseau et al. [16] estimated embodied GHG emissions in the global road material stock to be 8.4 $GtCO_2$-eq (lower estimate of 5.3 $GtCO_2$-eq, and upper estimate of 12 $GtCO_2$-eq) if the roads are built anew using current material production methods. This large ranges in estimates demonstrates the large variance in existing estimates of material stock and embodied emissions of stocks, which is a result of lack of data. In addition, road construction and maintenance are expected to increase in the future, as a considerable share of the global population still lacks access to basic road infrastructure [6]. Therefore, material demands and the associated embodied emissions from road construction can be expected to rise as well. Despite this, the challenges involved in limiting material demand and GHG emissions associated with road construction have received less attention in the literature than have the

challenges linked to buildings [7]. This lack of attention can at least be partially explained by the general lack of road-related data.

To counter such low data availability, the Global Roads Inventory Project (GRIP) has gathered and harmonized information related to road length and type of roads for 222 countries [14]. Rousseau et al. [5] have reported that in the GRIP dataset, more than 20% of the road length data is missing in many countries, while Central and South American and European countries have the lowest levels of missing data. Similarly, OpenStreetMap, which is another global-level dataset describing roads, suffers from data incompleteness [15]. In Europe, many gaps exist in the official Eurostat statistics on transport infrastructure [16]. In addition, non-government-owned roads, such as communal roads, have overall lower data quality and are sometimes not included in the national statistics [17]. The lack of data is a major challenge for the expansion of MSFA studies of transport infrastructures such as roads [12][13].

In general, MSFA approaches adopted to develop building stock models are categorized between top-down or bottom-up [17], [18]. Top-down models do not take into account the differences between individual buildings. For building material stock models, the top-down approach uses aggregated data on material use (e.g., material trade, consumption statistics) and applies lifetime estimates to derive the quantities of material accumulated in the building stock and material outflows [19]. In contrast, the bottom-up approach models a building stock "piece by piece", e.g., at the building level. This type of modeling is particularly important in decarbonization research because bottom-up approaches can be integrated with geospatial data in Geographic Information System (GIS), so as to gain a better understanding of the physical composition of the building stock [7]. Material stock models take into account those building characteristics that may impact their material content, such as age, use, and/or construction type. Regardless of their focus, modeling building stocks using a bottom-up approach requires substantial amounts of data [20] [7]. There are currently some international efforts into generating and collecting building inventory datasets, which includes varying degrees of completeness for the core attributes required for building stock modeling.

The most useful and relevant attributes are building footprints, heights, floor space, building type, and construction year [21]. However, most national-level inventory datasets do not contain comprehensive information on these attributes. For example, a harmonized European building inventory dataset using governmental data and OpenStreetMap has shown that for building height, construction year, and building type, only 73%, 24%, and 46%, respectively,

of the data are available [21]. In Europe, only Spain, The Netherlands and France have national-level databases that contain building footprints and attributes [21]. Two tools have been developed by the EU to collect building inventory data, each with its own limitations. The Copernicus Reference Data Access (CORDA) node only contains data from selected countries that do not include Sweden [22]. In contrast, the EU Building Stock Observatory provides data at the country level for all of the EU Member States, but only in an aggregated fashion [23]. The Microsoft Corporation has developed an open-source building footprint dataset that uses a combination of neural networks and aerial images, but building height coverage in the dataset is limited to US, Canada, Australia and countries in continental Europe but does not include Sweden [24]. This lack of data is a key obstacle to developing high-resolution building stock models.

2.1.2 Key takeaways

To summarize, data availability is a challenge for both the material stock and flow modeling of roads and residential buildings. An emerging approach to generating new data or impute missing data is machine learning.

2.2 Machine learning to impute missing inventory data

This section provides an overview of the current application of machine learning (ML) to generate or impute built environment inventory data and the associated challenges. Existing ML studies that predict an attribute of the built environment (e.g., age of building, or width of roads) are often focused on smaller geographical scales like cities and neighborhoods to leverage higher data availability at the expense of applicability (**Paper I** and **II**). The consequence of such an approach is that the developed model pipelines cannot be easily scaled up to larger geographical scales. This lack of scalability of ML approaches highlights the need to balance the aim for accuracy and the need for more generalizability. To achieve the goal of balancing accuracy and generalizability, this work develops machine learning models for both roads and buildings that achieve good accuracy without using scarce attributes as features.

2.2.1 Machine learning model for road width prediction

This subsection provides a background on the application of ML to predict road attributes. To understand how to develop ML models to predict road attributes, it is important to first understand existing works that utilize similar approaches. Several recent studies have applied

ML approaches to estimate the material stocks and flows of roads by predicting various types of road attributes [10], [32], [33]. Zhang et al. [33] employed a set of time series analysis-based ML models to project the historical material stock of Japanese roads from 2020 to 2050 under five different national shared socio-economic pathways. The aim of this study is to estimate future development of road stock at an aggregated level (prefecture) and thus do not use ML to predict specific road attributes. This strand of research that uses ML for forecasting of stock is not relevant for the aim of this study.

Another strand of research aims to overcome these limitations by using ML models to predict the depth of road layers. Ebrahimi et al. [10] estimated and predicted the material stocks and flows of the Norwegian road network by predicting the depth of roads with a decision tree-based ML algorithm. The strengths of this method are its abilities to incorporate the effect of traffic flows and to estimate the dissipative flows of materials. As the ML training process requires extensive data, the analyses were limited to national roads, for which all the input data were complete. Similarly, Wang et al. [30] estimated the material stocks and flows of road infrastructures in Belgium using a combination of ML models and the archetype-based approach. This work still requires road thickness data which is not widely available. These approaches can be seen as building upon and scaling up scarcely available data, but they do not address the lack of more fundamental attributes for the purpose of estimating material stocks.

While the abovementioned approaches advance the estimation of material stock and flows of roads, they do not fully address the fundamental challenge of missing road attribute data. Even within a country, the quality and availability of the data on road attributes can be highly heterogeneous (see Wang et al. [30]), and this data heterogeneity impedes the implementation of bottom-up MSFA studies of roads at the national or international level. Furthermore, the material stocks and flows of non-government-owned roads are often underestimated due to incomplete data [17]. A key data-point for estimating the material stock and flows of roads is the road width, to which material stock of roads is highly sensitive [34].

2.2.2 Machine learning model for building attributes prediction

This subsection provides an overview for the application of ML to predict building attributes. For buildings, machine learning-based approaches have also emerged as a method to fill in the gaps in incomplete datasets so as to achieve higher spatial resolution and accuracy [25]. Machine learning represents algorithms that can learn from the attributes of an input dataset, to generate a prediction based on the learning process (for an introduction to ML, see [26]). In

the context of building stock modeling, ML approaches are often used to predict building attributes such as age or height to complete or enrich building inventory datasets. Existing work using ML approaches for building inventory development have been applied at various spatial scales. Most of the existing work was performed at the urban scale (neighborhood or city), with some studies carried out at the national scale [27] [9] [28].

The limited number of studies conducted on the national scale is due to the high data requirements of existing ML approaches, especially the heavy emphasis on height data (See **Paper II**). As demonstrated previously by Arehart et al. [9], ML models are capable of providing good predictions of usable floor space if height data is partially available. Arehart et al. [9] uses building height data for the US to train a ML model that predicts height of the North American building stock and then use the predicted height to calculate usable floor space of buildings. At the time of writing, this is the only study that uses ML to predict usable floor space and only doing so indirectly.

As for building age, one noteworthy insight is that no previous studies predicted building age at the national level without the usage of building height data. Nachtigall et al. [28] predicted the age of residential buildings in the Netherlands, France, and Spain, using mainly urban form features with a mix of real height data and previously predicted height data from Milojevic-Dupont et al. [27]. Urban form features are ML training features calculated using building footprints and surrounding roads and building blocks without the need for height data (See **Paper II**). The review on previous work indicates that urban form features are good input features to train ML models and thus could be used to balance the need for accuracy and scalability. However, existing literature does not explore the possibility of using only urban form features without height data to predict building attributes.

2.2.3 Key takeaways

To summarize, there is a large need for data for more detailed MSFA studies and ML is a proven methodology to generate or impute data, but most existing approaches lack generalizability due to the reliance on scarcely available data such as building height and road layer thickness. Therefore, there is a need to develop a methodological approach that can be used to impute or generate built environment attribute data at a large geographical scale using widely available data.

2.3 Dynamic material flow analysis to model renovation

Dynamic material flow analysis (DMFA) of buildings quantifies the material flows from producing, operating, maintaining, and disposing of the stock [29]. Maintenance and renovation of buildings is necessary to maintain its function and aesthetics and to improve its energy efficiency, but existing modeling of maintenance/renovation in MSFA studies has been conducted with broad assumptions (see **Paper III**). The main imprecise assumption is that buildings are often modeled as a single product instead of a system of parts. This assumption overlooks and underestimates not only the material flows but also the associated impact such as embodied carbon emissions.

2.3.1 Existing studies on renovation

This subsection provides an overview of existing work on renovation of buildings. Similar to the literature on ML applications on built environment stock, studies that have modeled renovation of buildings have been focused on smaller geographical scales. In addition, very few studies have modeled both the material flows from renovation and the embodied carbon impacts. To the author's knowledge, the only study that estimates both the material flows from all renovation activities and the embodied emissions are by Ohms et al. [30], this study limits the geographical scope to a university campus. The rationale for limiting the geographical scope is due to data availability. Göswein et al. [31] developed a building stock model for a neighborhood in Lisbon, Portugal, to investigate the embodied emissions of renovation activities focusing exclusively on insulation materials. At the national level, Berrill et al. 2021 [32] developed a housing stock model with a focus on representing vacancy rates for each state in the USA. Subsequently, Berrill et al. 2022 [33] further developed the model to include renovation for only building envelopes. The material requirement for envelope renovation is assumed to be a percentage of material requirements for new construction. At the time of writing, no study has estimated the material flow of all renovation activities and its associated carbon emissions at a national level. This gap in literature can be partially explained by the challenge of modeling the dynamic interactions between the differing lifetimes of various 'parts' of the building.

2.3.2 Theorizing renovations with building shearing layers

The challenge can be attributed to a commonality to most DMFA studies is that they treat a building as a single product and thus a single lifetime. Such assumption is in strong opposition

with the architectural concept of building shearing layer, that has become increasingly central in circular construction literature, and which states that “there isn't any such thing as a building. A building properly conceived is several layers of longevity of built components.” (Duffy, quoted in Brand, 1994) Indeed, the vast majority of buildings are not monolithic; rather, they are composed of multiple parts, each with their own lifetime [34]. Brand et al. [34] theorized six shearing layers in a building, namely its site, structure, skin, space, services, and stuff (See fig.1).

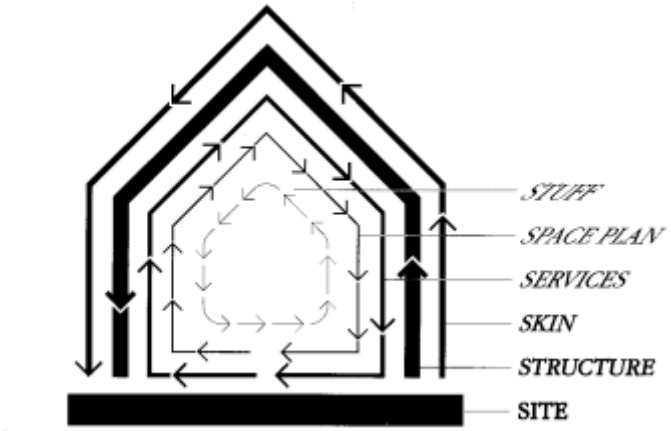

Figure 1. Illustration of the shearing layers for a building [34]**.**

When it comes to material stock (MS) and DMFA modeling of the built environment, three of the six shearing layers are particularly relevant: the structure, the skin (i.e., the building envelope), and the space layer. The latter refers to the building elements that divide the space within the building and that are not part of the structure, such as partition walls, ceilings, or flooring. Each shearing layer has a specific lifetime, making their differentiation particularly critical when looking at renovation dynamics in a building stock. Indeed, buildings typically undergo multiple renovations over their lifetime, with each renovation focusing on a specific layer [34]. For example, renovation of the façade of a building (skin layer) does not always happen at the same time as changing the interior layout of a building (space layer). To accurately capture the complex dynamics of building renovation activities, it is therefore essential to model material stock and flow at a higher level of granularity, so the various longevities of shearing layers are represented. In other words, dividing a building into its shearing layers can make building DMFA models more representative of real-world conditions.

2.3.3 Key takeaways

To summarize, the existing DMFA models of building renovations are coarse and use simplified assumptions. One such assumption is that most existing studies treat buildings as a single product with a single lifetime. This assumption of using a single lifetime fails to capture the dynamic interaction between different 'parts' of a building. By adopting the shearing layer concept to model the MS of buildings, different lifetimes can be assigned to different shearing layers in DMFA models. Therefore, there is a need to further develop existing DMFA models to be able to utilize the different lifetimes and capture interactions between different layers such renovation of the skin layer might not happen for a building if said building is too close to demolition.

2.4 Sweden as a case study

Sweden is chosen as a case study for its data availability. Indeed, Sweden is relevant due to its commitment to net zero emissions by the year 2045 [35]. Furthermore, the Swedish Transport Administration has committed to a goal of all state-owned infrastructure becoming climate neutral by 2040 [36]. For buildings, Sweden has committed to the EU's renovation wave policy to increase renovation rates for the purpose of energy efficiency [37]. This commitment to reduce emissions from the built environment makes Sweden particularly relevant in the context of studying material stock and flow modeling of the built environment and how to reduce the embodied emissions.

In addition, Sweden has relatively good data availability at the national level compared to other countries, but still incomplete. For roads, the Swedish Transport Administration has a comprehensive open-source dataset on state-owned roads with length data as well as other attributes such as archetypes [38]. This dataset however does not contain width data for all road segments, especially for municipally and privately owned roads (See **Paper I**). Widths are essential for estimating material stock of roads in a bottom-up approach, but in total 36% of road segments do not have width data.

For buildings, the Swedish Land Survey (Lantmäteriet) collects a building registry dataset that contains building attributes, mainly for taxation purposes. The attributes in the building registry dataset that is useful for building stock modeling are usable floor space, construction year of building, and building type. The usable floor space and construction year data are, however, incomplete (See **Paper III**). For single-family (SF) buildings, 16% are missing construction

years, and 13.4% are missing usable floor space data. The missing data problem is even more severe for MF buildings, with 19.9% missing construction years, and 62.7% missing usable floor space data. Furthermore, this building registry does not contain height data, which means height cannot be used as a proxy to estimate usable floor space.

3 Method

This section describes the main methods used to answer the research questions: *Machine learning applications* (Section 3.1), *Dynamic material stock and flow analysis* (Section 3.2), and *Scenario analysis for embodied carbon emissions reduction* (Section 3.3).

3.1 Machine learning applications

Machine learning or statistical learning represents algorithms that can learn from the attributes of an input dataset, to generate a prediction based on the learning process (for an introduction to ML, see Hastie et al. [26]). The two machine learning models used in this work are regression and classification models. Regression models are more suitable for prediction tasks with targets that are continuous variables such as widths of a road section **(Paper I)** or usable floor space of a building **(Paper III)**. Classification models on the other hand are more suitable for prediction tasks with targets that are discrete variables such as construction year of buildings **(Paper III)**. This subsection aims to provide an overview of the ML workflow in this study.

3.1.1 Dataset inspection and preprocessing

The first step of any ML workflow is to inspect and preprocess the dataset. This step includes understanding the dataset, such as the statistical distribution, completeness of data, the potential existence of outliers, and the like. The goal of the preprocessing step is to gain an initial understanding of the data and to decide on questions such as what type of features might be needed, and whether regression or classification algorithms are more suitable. The preprocessing step could also include removing outliers if they exist or data stratification.

One of the issues identified during the data inspection process pertained to class imbalance. A ‘class’ in a classification problem is one of the possible categories or labels that an input data can be assigned to by the model. Class imbalance is a problem for classification models when the underlying data are skewed and certain classes are under-represented, making the ML model’s ability to predict minority classes less-effective [39]. In the building dataset, a significant class imbalance existed for construction years of buildings, more precisely between 1960 and 1980. This observed imbalance of construction years can be attributed the Swedish ‘Million Program’ in which more than 1 million buildings were built in a ten-year span (1965 to 1974) [40].

3.1.2 Feature engineering

The second step of the ML workflow after inspecting and preprocessing the dataset is feature engineering. In ML, a feature serves as a numeric representation of a specific aspect of the raw data. Since ML models require an adequate number of features to capture and reflect effectively the underlying data's characteristics, feature engineering is essential, so it is the second step in the ML workflow [41]. This process involves extracting features from the raw data and converting them into formats that are suitable for utilization by the ML model [42].

As previously discussed in **Section 2.2**, existing ML models rely heavily on data with limited availability such as building height. To overcome this challenge, this work calculates and uses urban form features as the main input in the ML workflow. The advantage of urban form features is that these features can be calculated using only building footprints and road data. The quantitative analysis of urban morphology distinguishes between three key building elements: footprints, plots, and street-based blocks [43], as illustrated in **Figure 3**, where tessellations enclosed by streets are used as a proxy for building plots. Each element captures the unique attributes of the corresponding building, and the features are generated based on the elements.

Building elements refer to the 2D geometry of the footprints of buildings, which contain useful information that can be used for predicting construction age [44] and floor space [9] through ML models.

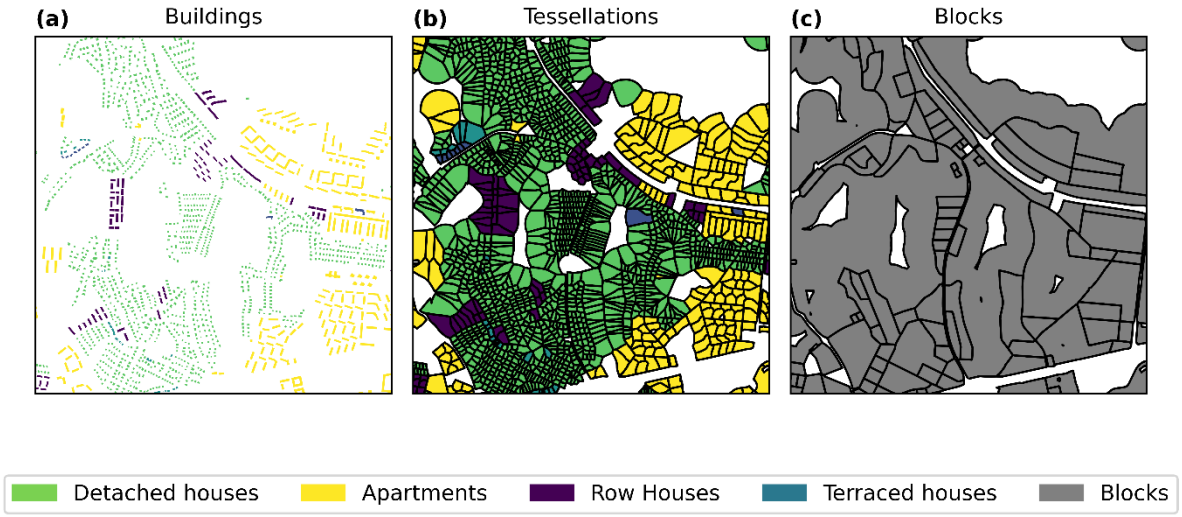

Figure 2. Illustration of the three elements of urban morphology using an area of Gothenburg (Sweden) as an example: (a) building footprints; (b) tessellations enclosed by streets, used as a proxy for building plots; and (c)

building blocks generated using tessellations. Note that no color coding is used here because the blocks include a variety of building uses. Source: **Paper II**.

For the ML predictions of usable floor space and construction year of residential buildings, 16 urban morphology features are generated and applied to the three elements of urban morphology for a total of 48 features. All features are calculated using the Momepy package developed in Python using a combination of the building's GIS footprint data and road network GIS data [45]. The same set of features are used for both the classification and regression model. For the ML prediction of width of roads, 12 road network features are calculated using only road GIS data using Momepy. In addition, social economic features such as population and building features such as distance to the nearest buildings are used as features.

3.1.3 Classification models training and validation

The next step in the ML workflow after generating features is the model training and validation process. Classification models are used to predict building construction years (**see Paper II**). As previously discussed in **Section 3.1.1**, there is a significant class imbalance of construction years. Existing ML algorithms can mitigate some of the class imbalance problem, but not always in a satisfactory manner. Therefore, an additional step is required to counteract the class imbalance and maximize the ML model's performance [46].

There are multiple available techniques to tackle the class imbalance problem, and the under-sampling technique is chosen due to the underlying distribution of the data (See **Paper II**). Therefore, the open-source Python package 'Imbalanced-learn' is used to test four under-sampling techniques to identify which technique can produce the optimal prediction performance [47]. The first technique is the Random Under-Sampler (RUS), which randomly under-samples the majority classes without replacement. The second is the Near Miss (NM) method, which selects a subset of the majority classes samples that are closest to the minority classes samples [48]. The third is the One-sided Selection (OSS) technique, which initially finds the observations that are hard to classify and then removes noisy samples [49]. The fourth is the Neighborhood Cleaning Rule (NCR) method, which uses a combination of edited-nearest-neighbor and a k-nearest-neighbor to remove noisy samples from the dataset [50]. Subsequently, we use the XGBoost algorithm for the classification task and train the model on the four different under-sampled datasets. XGBoost is chosen as it has been successfully applied to predict various attributes of the building stock with high levels of precision [28],

[51]. Therefore, we test the XGBoost algorithm with different resampling methods to address class imbalance issues in the dataset.

To evaluate the result from the under-sampling techniques, a set of evaluation metrics are used. The metrics used in the study and their associated equations and descriptions are shown in **Table 3** in **Paper II**. There is currently no clear consensus in literature on what are the most accurate evaluation metrics, so we use the most popular metrics. The best performing under-sampling technique is then used to train the final classification model that predicts the construction year of buildings.

3.1.4 Regression models training and validation

The road width prediction problem in **Paper I** and the usable floor space prediction problem in **Paper II** are tackled with regression ML models. Since regression models do not face class imbalance problems, we compare the four most widely used algorithms to test which has the best performance instead. One of these is XGBoost [52], which is also used for the classification model. XGBoost is an ML algorithm that is a member of the family of gradient-boosting techniques [55]. XGBoost is an ensemble learning method that combines the predictions of multiple individual decision trees, known as weaker learners, to create a strong final predictive model. In gradient boosting, new models are built sequentially by correcting the mistakes of the previous model. Each new model is trained to predict the residual errors of the ensemble of previous models.

In addition to XGBoost, we test three ML algorithms that are more commonly used for regression problems. Random forest (RF) is an ensemble learning algorithm that builds a collection of decision trees and combines their predictions to create a more-accurate and robust model [53]. The RF algorithm is implemented using the scikit-learn package in Python [54]. LightGBM is another gradient-boosting decision tree algorithm developed by Microsoft that has higher training efficiency and lower memory usage [55]. CatBoost is yet another gradient-boosting algorithm developed to offer more support for categorical features [56]. For both the regression and classification models, a five-fold cross-validation method is used to prevent the model from overfitting to the training data. Overfitting refers to an ML model that learns the training data too well, including the noise to the point that it performs poorly on test data. The hyper-parameters of the models are optimized using the Optuna Python package [57].

To validate the results of the ML models, several evaluation metrics are used. The regression models use a set of more traditional evaluation metrics. The first metric is the Mean squared error (MSE) as the evaluation metric during the model training process. MSE measures the average squared differences between the predicted and actual values and is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

where n is the total number of data-points, y_i is the actual value of the i-th observation, and $\hat{y}_i$ is the predicted value of the i-th observation.

Two additional evaluation metrics are used to compare the different ML algorithms for regressions: Mean absolute error (MAE), and R-squared score (R^2). MAE measures the average absolute differences between the predicted and actual values, defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

where n is the total number of data-points, y_i is the actual value of the i-th observation, and $\hat{y}_i$ is the predicted value of the i-th observation. MAE penalizes errors to a lesser degree than the MSE, but it is a more-intuitive evaluation metric.

The R-squared score measures the proportion of the variance in the dependent variable that is predicted from the independent variables and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where n is the total number of data-points, y_i is the actual value of the i-th observation, $\hat{y}_i$ is the predicted value of the i-th observation, and $\bar{y}$ is the mean of the target variable.

Like the classification model training and validation process, the best performing regression algorithm is chosen and then used to predict the road width and usable floor space.

3.2 Material stock and flow analysis

This section provides an overview of the methods and equations used for the modeling of the material stocks and flows of the Swedish roads and residential buildings. Equations 4-6 are generalized for both roads and residential buildings, and specific equations can be found in the appended papers.

3.2.1 Scope of study

This subsection provides an overview of the scopes of the papers appended in this summary. **Paper I**, which focuses on roads, includes all paved roads and gravel roads, but excludes dedicated bike lanes and pedestrian walkways. The road layers included in the material intensities are surface course, binder course, base, sub-base embankment (as defined in Lanau et al. [58], see Fig (1)). We do not distinguish between above ground and underground in this study. The materials included in the MI are asphalt, steel, and aggregates. Concrete roads are excluded from the system boundary, as there are only 68 km of concrete roads in Sweden as of the year 2022 [63].

For buildings studied in **Paper III**, the scope of study includes all single-family (SF) buildings and multi-family (MF) buildings. The SF buildings are assumed to be one structure type across 12 age cohorts (each cohort is 10 years, from 1880 to 2000). The predominant structure type of SF buildings is assumed to be timber for all age cohorts. The MF buildings are assumed to include four structure types, and 13 age cohorts (each cohort is 10 years, from 1880 to 2010). The structure types include Wood multi-family (WMF) buildings, Wood-brick multi-family (WBMF) buildings, Brick multi-family (BMF) buildings, and Concrete multi-family (CMF) buildings. The data on building structure was retrieved from the study of (Bian et al, in prep) who used building morphological indicators to predict building structures of residential buildings in Gothenburg. Results showed high accuracy (more than 80%), and we use them in this study to be able to apply Swedish MIs (which follow a structure-use-age archetype) to the inventory. The material intensity includes 12 materials for both SF and MF buildings (See **Paper III**). Furthermore, the scope of the MI dataset for both SF and MF buildings includes the superstructure (above ground), substructure (underground), and foundation's compact layers.

The geographical scope of both **Paper I** and **Paper III** is Sweden. The temporal scope of **Paper I** is from 2022 to 2045 while the temporal scope of **Paper III** is from 2025 to 2050. The DMFA in **Paper I** is conducted at the regional scale (Sweden divided into 6 regions based on different lifetime estimates) and the DMFA in **Paper III** is conducted at the municipal level (290 municipalities in Sweden). **Paper II** is focused on method development and the geographical scope is Sweden.

3.2.2 Building renovations

The following two subsections provide an overview of the logic and assumptions behind the use of DMFA model to estimate material flows from building renovations and road maintenance. Both DMFA models are coded in Python using the Open Dynamic Material Systems Model (ODYM) [59].

We model the material flows from buildings using consecutive bottom-up stock-driven DMFA models and implemented in the following order: a stock-driven model that quantifies the future stock and outflows of usable floor space based on age cohorts. The outflow result from the first DMFA model is the amount of floor space demolished each year. To determine the inflows and outflows associated with the renovation activities, the future stock is used in another stock-driven model using lifetimes specific to the skin layer to simulate the need for renovations.

In the second DMFA model, renovation activities of buildings are modelled by assigning a specific lifetime to the skin and space layer of buildings. The surviving stock $S_{survive_{i,c}}(t)$ (see Equation (5) below) from the first DMFA model is then used as the stock that will undergo renovation activities. In other words, the main logic behind the modeling of renovation activities is that only the surviving stock from each age-cohort is renovated. These two consecutive DMFA models are an adaptation of the convolution approach introduced in Sartori et al. [60]. This method is named the *layered model* due to the use of convolution of lifetime between the building shearing layers. In addition, we compare the results from the layered model with the stock-driven model that treats the building as a single object, which we call the *monolithic model*. In the monolithic model, instead of using the surviving stock to conduct a second consecutive DMFA model, parallel stock-driven models to the first DMFA model are conducted to represent renovation activities.

3.2.3 Road maintenance

A key assumption behind the modeling of road maintenance in **Paper I** is that no roads are demolished in Sweden. Therefore, only one stock-driven DMFA is conducted to estimate the material flow from the maintenance of roads. In addition, it is assumed that only 50mm of asphalt is removed and repaved from maintenance according to expert opinion (from PEAB construction company that works with road maintenance in Sweden).

3.2.4 General stock and flow modeling approach

The first step is to calculate the in-use material stock, and the stock is calculated by multiplying the inventory with material intensity, as shown in Equation (4):

$$MS_i(t) = MI_i(t) * Inv_i(t) \quad (4)$$

where $MS_i(t)$ is the material stock of built environment type i at time t in tons, $MI_i(t)$ is the material intensity of built environment type i at time t in tons, and $Inv_i(t)$ is the inventory of built environment type i at time t. The inventory used in **Paper I** is road surface area in m^2 (length multiplied by width), with both existing and ML-predicted width data. The inventory used in **Paper III** is the usable floor space of buildings in m^2. The inventory is the result of **Paper II**.

The first step of the stock-driven DMFA model is to calculate the outflow from existing stock, as shown in equation (5):

$$outflow_{i,c}(t) = \sum_{\tau=t_0}^{t} inflow_i(\tau) * (1 - Survival_C \quad (t-\tau)) \quad (5)$$

where $outflow_{i,c}(t)$ is the outflows or demolition of stock type i from age-cohort c (construction year) at the end-of-life (EoL) at time t in tons, and $Survival_c(t-\tau)$ is the survival function that represent the share of the inflow from age-cohort c remaining in the stock at time t. The survival table is constructed from an age-cohort dependent lifetime distribution. There is no consensus in literature on functional forms of lifetime distributions [61]. The Weibull distribution is chosen as it is the most commonly used and most available data exists [62].

The next step is to calculate the inflow of building type *i* at time *t* using the mass-balance principle, as shown in the following equations:

$$S_{survive_{i,c}}(t) = \sum_{\tau=t_0}^{t} (inflow_i(\tau) - outflow_{i,c}(\tau))$$

$$inflow_i(t) = S_i(t) - S_{survive_{i,c}}(t) \quad (6)$$

where $S_i(t)$ is the stock type *i* at time *t* in tons and $S_{survive_{i,c}}(t)$ is remaining or surviving stock type *i* at time *t* in tons after $outflow_{i,c}(t)$ is removed.

3.3 Embodied emissions

The embodied emissions from the construction and maintenance of the built environment stocks are calculated using Equation (7):

$$Embodied_{emission_t} = \sum_{i,t,m} M_{inflow_{total_{i,t,m}}} * EF_m \quad (7)$$

where $Embodied\_emission_t$ is the total embodied CO_2 emissions in year t in tons, $M\_inflow\_total_{t,m}$ is the total inflow of material m from stock type i at time t in tons, and EF_m is the emission factor for material *m* in tons/ton. The emission factors (*EF*) used here are based on estimates made by Karlsson et al. [68]. The emission factors (EFs) used in this study account only for life cycle stages A1–A3[1], thereby excluding emissions associated with later stages such as transportation to site and construction (A4–A5).

[1] Stage A1 refers to raw material extraction, Stage A2 refers to transport to manufacturing site, and stage A3 refers to manufacturing.

4 Selected results

This section presents some selected results from the three appended papers. Selected results are shown in four main areas: ML imputation, material stock, material flows with a focus on renovation and maintenance, and embodied carbon emissions.

4.1 Machine learning imputation

The results of the regression ML models that predict road widths (**Paper I**) and residential building usable floor space (**Paper II**) are presented in **Table 2**.

Table 2. Results of the evaluation metrics for each regression machine learning algorithm. For MAE, the lower the absolute value the better the performance, and for R^2 the closer to 1 the better the results.

	Road width prediction		Building floor space prediction	
ML models	MAE (m)	R^2	MAE (m^2)	R^2
XGBoost	0.567	0.784	28.74	0.789
LightGBM	0.576	0.781	29.43	0.788
CatBoost	0.669	0.728	31.19	0.781
Random Forest	0.623	0.748	32.16	0.756

MAE, Mean absolute error

The best-performing model across both studies is XGBoost, achieving similar results. The hypothesis behind the similar R^2 result is that a R^2 value of close to 0.8 is the current upper bound of gradient-boosting ML algorithms without overfitting. The block-box nature of ML algorithms however means that it is currently not possible to pinpoint the exact reasons for the performances.

The results of the classification ML model that predicts residential building age are shown in **Table 3**. The evaluation metrics used are Area Under the Precision-Recall Curve (AUPRC), Mean Area Under the Curve (MAUC), G-means, and Mean Mathews Correlation Coefficient (MMCC). The higher the score for each metric, the better the performance for the under-sampling method.

Table 3. Results of the evaluation metrics for each under-sampling method. For each of the evaluation metrics, a higher score (closer to 1) represents a better prediction performance.

Under-sampling method	AUPRC	MAUC	G-mean	MMCC
RUS	0.631	0.783	0.532	0.243
Near Miss	0.663	0.922	0.760	0.564
OSS	0.671	0.912	0.777	0.571
NCR	0.823	0.974	0.911	0.814

AUPRC, Area Under the Precision-Recall Curve; MAUC, Mean Area Under the Curve; MMCC, Mean Mathews Correlation Coefficient; RUS, Random Under-sampler; OSS, One-sided Selection; NCR, Neighborhood Cleaning Rule.

The results show that the choice of under-sampling method has a relatively potent impact on the performance level of the classification model. The difference between the MMCCs of the best-performing NCR under-sampling method and the worst-performing sampling method is 70.1%. The reason for this disparity in performance level is likely due to the random nature of RUS in removing valuable and informative data points from the sample. Based on all the evaluation metrics, the NCR produces the best results.

4.2 Material stock

The results presented in this section are synthesized from **Paper I** and **Paper III**. The material stock of roads and residential buildings for each county in Sweden is shown in Figure 3. In total, the material stock accumulated in roads in Sweden amounts to 1,950 Mt and the material stock accumulated in residential buildings to 267 Mt.

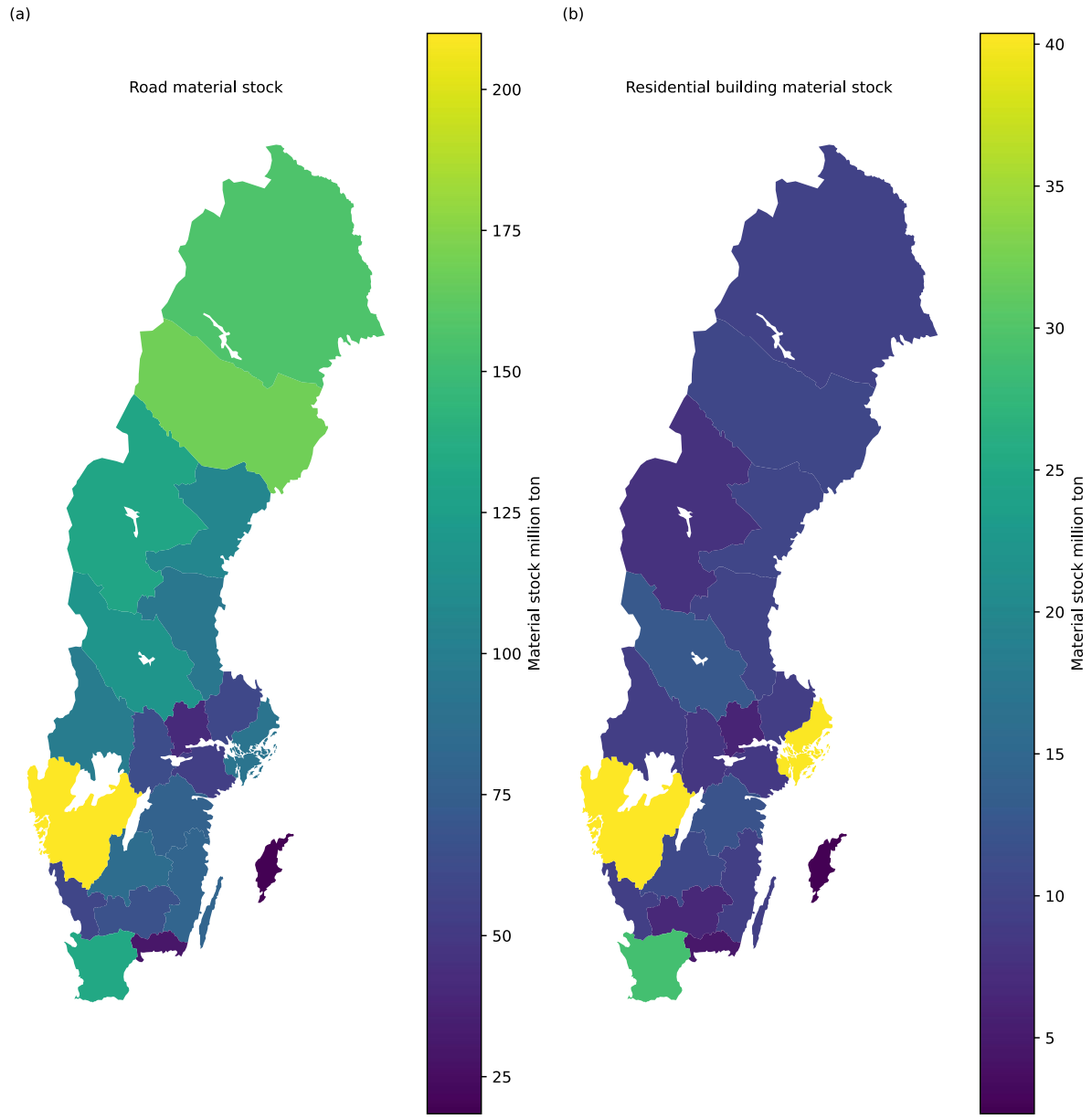

Figure 3. Material stock accumulated in roads and residential buildings for each county in Sweden, the unit is million tons (10^6).

The county with the highest accumulated absolute amount of total material stock accumulated in roads is Västra Götalands where Gothenburg, the second largest city in Sweden is located. The counties with the second and third largest amount of material stock accumulated are Västerbottens and Norrbottens, which are both counties in the far north of Sweden with vast geographical areas and thus the need for more roads. Stockholm as the largest city in Sweden does not have the most amount of absolute material stocks accumulated in roads due to the relatively small geographical area of the county.

The counties with the most material stock accumulated in buildings are Västra Götalands, Stockholm, and Skåne, where the three largest cities in Sweden (Gothenburg, Stockholm, and Malmö) are located. Västra Götalands county has 0.26 Mt more material stocks in residential buildings compared to Stockholm, which is due to the county of Västra Götalands encompassing a much larger rural area than the county of Stockholm.

Due to the large geographical area and sparsely distributed population, absolute material stock does not show the full picture of the distribution and accumulation of material stock. Therefore, the material stock per square meter for each county in Sweden is shown in Figure 4.

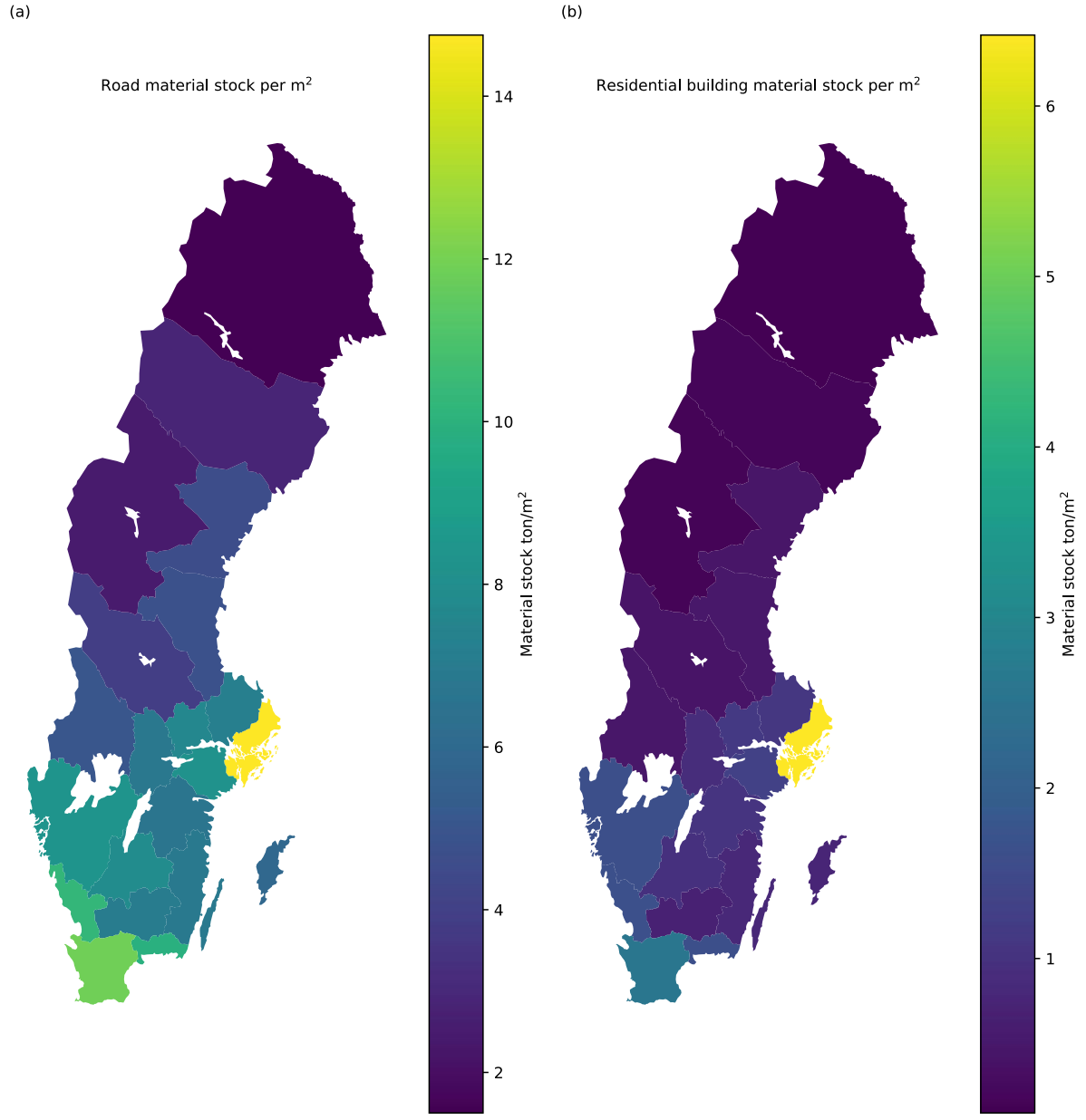

Figure 4. Material stock accumulated in roads and residential buildings for each county in Sweden, in t/m^2.

Normalizing material stock by county area reveals more clearly that accumulation is primarily concentrated in population centers. Moreover, there is a clear North-South divide, where most of the Swedish population resides in the southern part of the country.

4.3 Material flows

The results presented in this section are also synthesized from **Paper I** and **Paper III**, with a focus on the results from the modeling of renovation activities from residential buildings.

4.3.1 Material inflows and outflows

The inflows and outflows for all materials in roads and selected materials in residential buildings are shown in Figure 5.

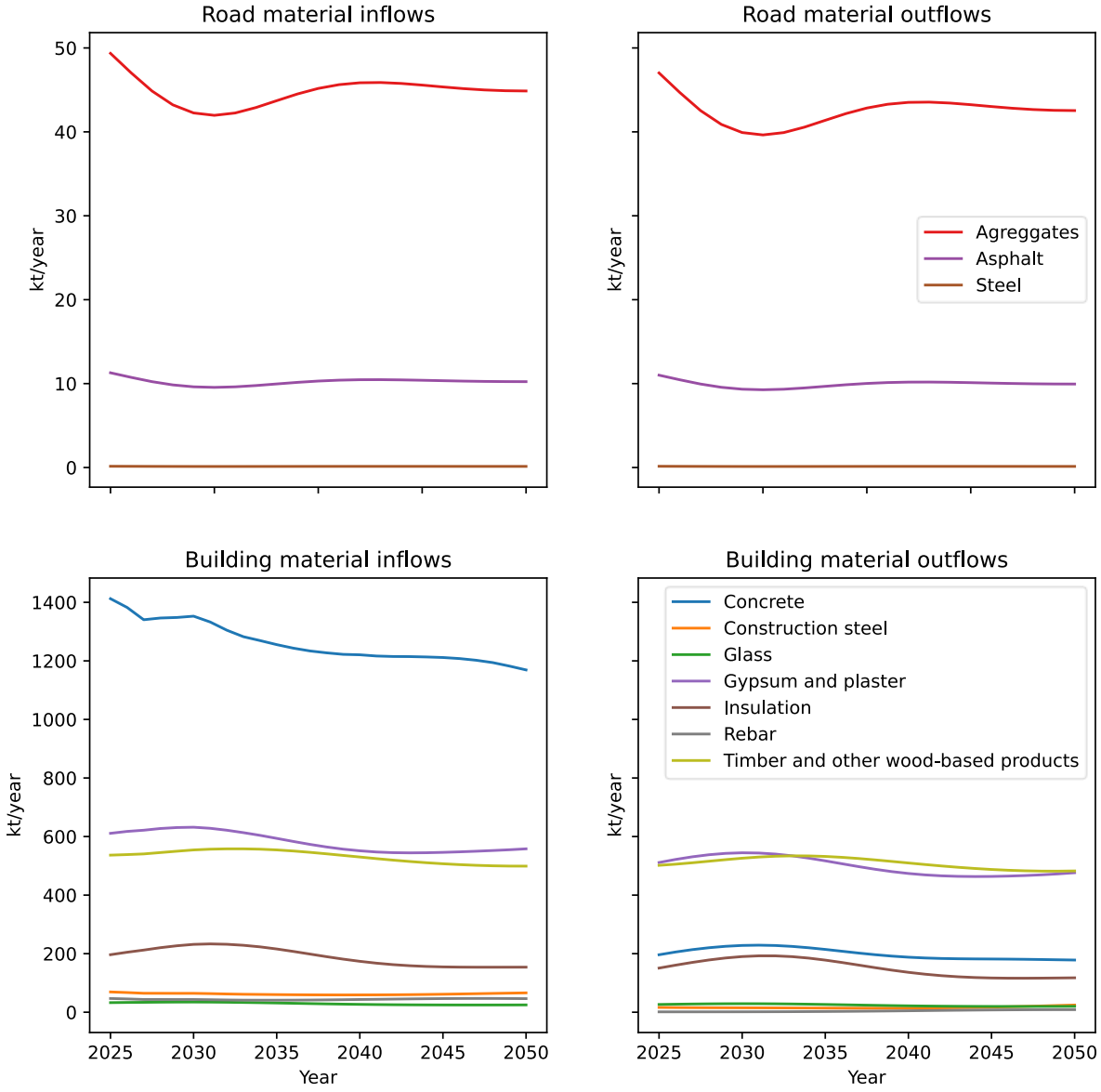

Figure 5. Material inflows and outflows from roads (top graphs) and residential buildings (bottom graphs) in Sweden from 2025 to 2050, in kt/year.

The result shows that the inflows and outflows of materials from roads are strikingly similar. The reason is that the difference between inflows and outflows is the new construction of roads. In the scenario used in the analysis, the new construction of roads is significantly lower than the amount of road segments that require maintenance, hence the small difference in inflows and outflows. Furthermore, we assume that the MI for roads are the same regardless of the time of construction, which makes the inflow and outflow more homogenous. This assumption of using the same MI is a limitation and could be improved in future studies.

The difference between material inflows and outflows of buildings is higher than the difference between inflows and outflows of roads. The material with the largest difference between inflow and outflow is concrete, which is due to the relatively low amount of demolition of residential buildings in Sweden. In addition, we assume that the sub-structure of buildings is not demolished. The outflow of timber and other wood-based products is expected to exceed the inflow, which means that there are high circularity potential for timber and other wood-based products. This higher circularity potential is primarily attributed to the prevalence of timber structures in buildings constructed before 1950, which are increasingly reaching the stages of renovation or eventual demolition. Furthermore, due to design optimizations, newly constructed buildings are expected to use less materials per square meter, leading to this difference in inflow and outflow for timber buildings.

4.3.2 Comparison between the layered and monolithic model

To better understand the results of the layered model for estimating renovation activities of residential buildings in **Paper III**, there is a need for comparison between the results of the layered model and the monolithic model.

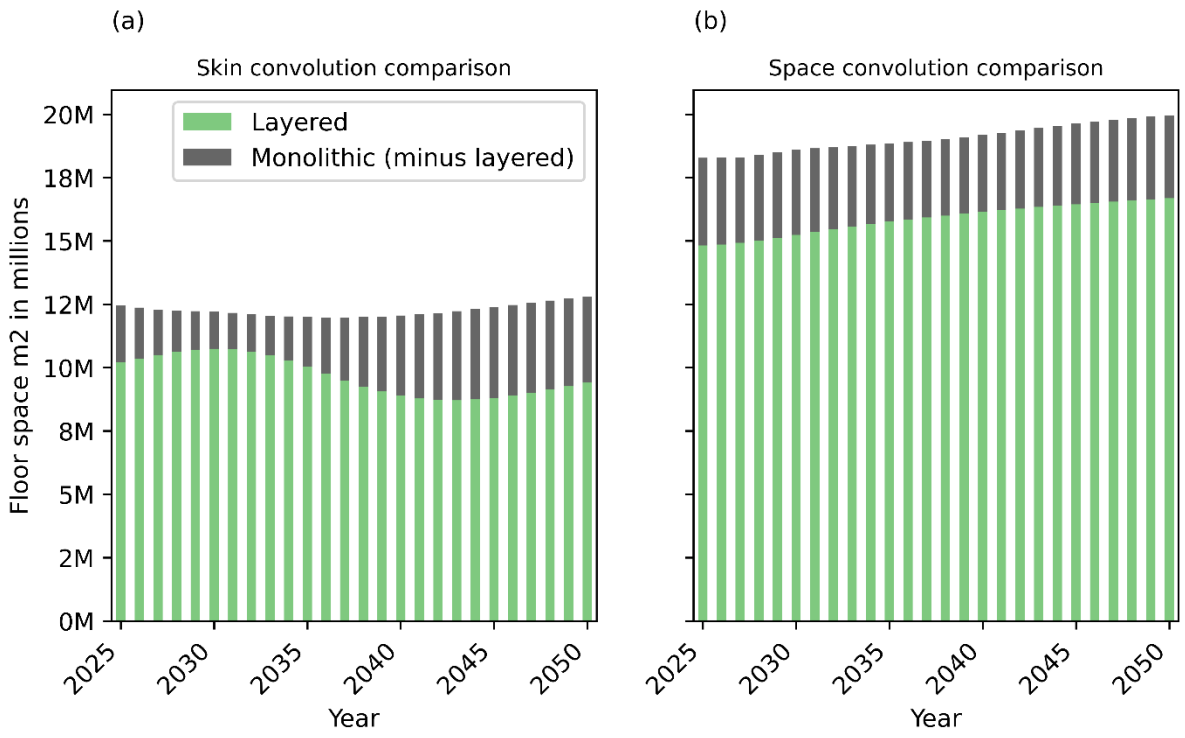

Figure 6. Comparison of the estimated amount of floor space that requires renovation from the layered and monolithic model for the skin and space layers. Green represents the results of the layered model. Grey represents the difference between the monolith layered models: the value hit by the grey bar corresponds to the results of the monolith model.

The comparison between the results from the monolithic model and layered model is shown in Figure 6. For both the skin and space layers, the monolithic model results in a higher estimation in terms of the amount of floor space that needs to be renovated. For the skin layer, the monolithic model yields an estimate that is, on average, 28% higher. For the space layer, the monolithic model yields an estimate that is, on average, 20% higher. This difference in estimation is a result from the monolithic model's assumption that the lifetimes of the skin and space layers are independent—an approach that leads to the renovation of building stocks nearing demolition.

4.4 Embodied emissions

The material flow results were used to estimate embodied emissions from maintenance and renovation activities. Those embodied emissions are the focus of this section, as they are deemed as the main findings of the study. They are shown in Figure 7.

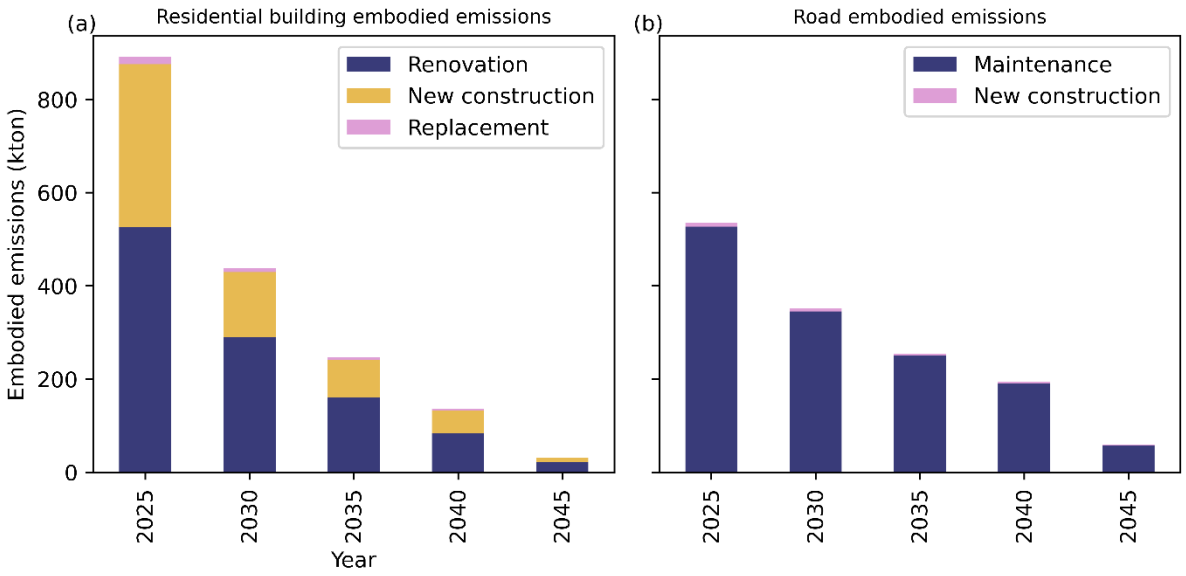

Figure 7. Embodied emissions from renovation/maintenance activities and new construction for (a) residential buildings and (b) roads. The replacement represents the replenishment of demolished building stock.

For both roads and residential buildings, maintenance and renovation activities are responsible for the largest share of embodied emissions. Since roads are almost never demolished and new construction is relatively limited, it is logical that maintenance of the top layer is the main contributor to embodied emissions. It should be noted that the embodied emission factors used for estimating the emissions from roads here have been updated compared to the values used in Paper I when it was published to incorporate updated estimates on emission factors.

In the case of buildings, the new construction scenario is relatively ambitious, as it assumes that all projected population growth will be accommodated by constructing new buildings, maintaining the current average floor space per capita for each municipality. This implies that, in practice, the volume of new construction is unlikely to exceed the scenario's assumptions, and therefore the proportion of embodied emissions attributable to new construction is also unlikely to be overestimated. Therefore, it can be safely assumed that renovation activities will be the largest contributor to embodied emissions from residential buildings. In addition, the renovation activities do not include deep energy efficiency renovations, which will be modeled in future work. The share of embodied emissions from renovations will be even higher if deep energy renovations are included, which requires more materials.

5 Conclusion and main findings

This section synthesizes the main findings of the study by addressing the research questions and the implication of the results. In addition, the contribution of this work to the field of research is discussed. Furthermore, the insights from the synthesis of results from the three appended papers are discussed. Lastly, the policy implications of the research are presented.

5.1 Main findings

This subsection is structured to sequentially answer the research questions in the order listed above (RQ1-3) and discuss the contribution of this work to the field of research.

5.1.1 Effectiveness of ML models

The results from **Paper I** and **Paper II** show ML models can be effectively used to predict missing attributes in the road and residential building inventory datasets. This proposed workflow can be especially relevant for larger geographical scales such as the national level where data such as building height might be missing or incomplete. The two main findings of this work with regards to the effectiveness of ML to impute missing inventory data at the national level (RQ1) are: 1) Urban form data are effective features for training ML models, 2) Classification models perform well for construction year predictions, 3) Regression models perform well for predictions of continuous variables such as road width and building usable floor space.

Due to the black-box nature of machine learning, it is not possible to conclude with certainty that any specific workflow will guarantee the best results. Therefore, the application of ML models to predict built environment attributes should be tested more on a case-by-case basis. The main learning from the process of training and testing three ML models as part of this work is that data inspection and preprocessing of data is very important. Questions such as the choice of using classification model or regression, and what type of preprocessing steps (resampling and/or outlier removal) are required depend heavily on the knowledge and understanding of the dataset. For example, ML models might not be effective if the proportion of missing data is too large. For some datasets and prediction targets, ML models might not be the most effective choice and normal regression models might be more appropriate.

Furthermore, the choice of features also depends on understanding what data could be predictive for the prediction target. For example, urban form features can be effective at

predicting physical attributes of buildings and roads such as road width and building usable floor space but might not be effective at predicting targets that are more related to user behavior such as energy consumption. Another learning from the ML model training process is that for this specific building inventory dataset, socio-economic data are not predictive features. Socio-economic data are not predictive and thus are not included in the final model so the unpredictive features do not add more noise to the model. In contrast, socio-economic data are predictive in **Paper I** for predicting road widths. Therefore, urban form data features should be tested on more datasets in the future to reach more generalized conclusions.

At a conceptual level, this study contributes to the broader research field by demonstrating that ML is a viable approach for researchers conducting bottom-up modeling of material stocks in areas where detailed inventory data are lacking. In the context of road material stock modeling, to the author's knowledge **Paper I** at the time of publication is the first work to demonstrate that urban form data can be effectively used to predict road widths. In the context of building material stock modeling, the main contribution of this work is to demonstrate that ML models can predict building attributes without height data. Height data of buildings are scarce (See **Paper II** for review), and this work demonstrates using urban form data are the main predictive features to bypass the use of height data. Urban form data have been demonstrated by Arehart et al. [9], and Nachtigall et al. [28] to be able to predict building height. To the author's knowledge at the time of writing, **Paper II** is the first work to use urban form data as input to a ML model direct predict usable floor space. Furthermore, to the author's knowledge at the time of writing **Paper II** is the first to use urban form data to predict construction years of buildings at the national level without using height data. It can be concluded based on the results from **Paper II** that ML models can be applied to more geographical areas where building heights are not available to enable more bottom-up material stock models.

5.1.2 How to improve DMFA modeling of renovation

This subsection focuses on the improvement of existing DMFA models to model renovation of buildings (**Paper III**). The research gap identified in DMFA modeling of renovation is that existing studies generally treat buildings as a single object with a single lifetime instead of a system of products with different lifetimes, which does not capture the reality that buildings have different 'parts' with different lifetimes. This work implements shearing-layers-based MI data when calculating material stock to better reflect reality. Furthermore, a workflow of two consecutive stock-driven DMFA models is developed to capture the interdependence of

different layers of a building. The main finding is that the monolithic model estimates total material flows from renovations to be approximately 20% higher than those estimated by the layered model. Since the layered model is introduced to bring the model closer to reality, the higher estimation from the monolithic model points to an overestimation of renovation activities.

While the layered model represents a step towards more accurately describing real world conditions accurately estimating material flows from renovation activities remains a significant challenge. This study assumes that buildings undergo renovation cycles that follow a normal distribution from the construction year of the building. This assumption is necessitated by the absence of reliable data on the timing of past renovation activities. Moreover, building renovations are not solely determined by physical lifespan but are also influenced by factors such as aesthetic preferences, economic considerations, and policy incentives. This shortcoming could be tackled by soft-linking the layered model to a model that can capture behaviors such as an Agent-based model (ABM) [63] . Furthermore, the appropriate lifetime parameters for modeling the renovation of the skin and space layers within a DMFA model remain an open research question. The built-environment stock and flow modeling community would benefit from more systematic studies on the lifetime of skin and space layers.

The main contribution of this work to the broader research field is to propose and demonstrate how shearing-layer-based MI can be leveraged to improve DMFA modeling of building renovations. To the author's knowledge, at the time of writing, this work is the first DMFA study to model renovations of buildings at the national level. This work represents a step forward to developing a modeling framework to conduct accurate bottom-up modeling of material flows from buildings at a large geographical scale. A further takeaway is that more work should be done to collect shearing-layer-based MI data to enable similar analysis in other geographical regions. The finding that the monolithic model may overestimate material flows from renovation activities warrants further investigation to determine whether this conclusion holds at broader spatial or temporal scales.

5.1.3 Embodied emissions from the Swedish built environment

The main finding from the embodied emissions calculation according to the embodied emissions scenarios is that neither the residential buildings nor the roads can reach net zero embodied emissions by the year 2045 absent additional mitigation measures. The results suggest that supply-side decarbonization within industry alone is insufficient to achieve net-

zero emissions targets, highlighting the need for complementary demand-side measures. From the perspective of the construction industry, measures such as design optimization—such as reducing concrete use—and enhancing material circularity through increased reuse of components and materials can play a critical role in supporting decarbonization efforts. Another set of potential measures that can support the decarbonization efforts are better utilization of the existing building stock. Such measures can include, for example converting under-utilized or unused offices or other non-residential buildings into apartments, adding an additional floor to existing apartments, and urban densification by building more new apartments in existing neighborhoods. Implementing these measures to improve building stock utilization will require supportive policies and potential regulatory changes to enable and incentivize such practices. Furthermore, greater attention could be given to measures that reduce the absolute demand for new housing, through so-called sufficiency policies. A key challenge associated with sufficiency policies is their potentially low level of social acceptance, which can hinder their implementation and effectiveness.

The estimation of embodied emissions in this study is subject to several limitations. The first limitation is that as previously mentioned, the emission factors only cover life cycle stages A1-A3. The exclusion of life cycle stages A4–A5 implies that the embodied emissions estimates presented here do not capture the full emissions profile of the construction sector supply chain—an important stakeholder and actor in broader decarbonization efforts. The embodied emissions from life cycle stages A4-A5 will be included in further studies to address this limitation. Another limitation of the estimation of embodied emissions is that all construction materials are assumed to be produced domestically. This assumption primarily impacts globally traded materials, such as steel, while its effect on locally produced materials like cement and concrete is less significant. In future studies, the uncertainties associated with these assumptions could be quantified through an uncertainty analysis. Finally, timber and other wood-based products are not treated as carbon-negative in this study (but climate neutral), which could be included in future studies.

The primary contribution of this study to the broader research field is the development and demonstration of a bottom-up workflow for estimating embodied emissions in the built environment, considering limited inventory data. To the best of the author's knowledge, this work is the first to provide an estimate of embodied emissions from Swedish residential buildings, including renovation activities. The workflow developed in this study can be adapted

and modified for use in other geographical regions with limited inventory data, thereby contributing to global decarbonization efforts.

5.2 Insights from the synthesis of the appended papers

The first additional insight is from the spatially-material stock comparisons shown in Figure 3 and Figure 4, which highlights the differences between road and residential buildings in-use material stocks. In absolute terms, the total in-use material stock of roads is about 7 times higher than the in-use material stock of residential buildings. However, most of the material stocks from roads are aggregates in the base and sub-base layers, which leads to the insight that compares embodied emissions from roads and residential buildings shown in Figure 6. Despite the higher absolute material stock in roads, the embodied emissions from residential buildings are higher than roads. Furthermore, the share of embodied emissions from maintenance and renovation activities are higher than new construction for both roads and residential buildings. The finding that renovation and maintenance activities make up the largest share of embodied emissions for roads and especially residential buildings is the key finding of this work, and the policy implications of this finding are discussed below (**Section 5.3)**.

The comparison of the machine learning models employed in **Paper I** and **Paper II** also yields additional insights. A key take-away is that urban form data are useful as features to predict both road and residential building attributes. Furthermore, for both the prediction of road width and building usable floor space, XGBoost is the best performing ML algorithm. Therefore, XGBoost should be considered in future work when training ML regression models.

5.3 Policy implications

This subsection provides discussions around the policy implications from this work. The main policy implication related to the overall findings in this work is that more attention needs to be paid to renovation and maintenance activities in the future with regards to reducing embodied emissions. The Swedish National Board of Housing, Building and Planning (Boverket) has proposed to the Swedish government to accelerate the introduction of limit values for climate impact from buildings to reduce emissions from buildings [64]. The Boverket highlights that embodied carbon emissions from renovations are a major share of the overall emissions but also indicates that there is currently limited amount of studies on the building stock level. The work in **Paper III** fills in this gap and provides support and confirmation that indeed embodied emissions from renovations, and more specifically not just from energy efficiency renovations

should be considered in limit values. The discourse on embodied emissions from renovations often focus on materials related to energy efficiency renovations such as windows and insulation materials [64]. However, the results from **Paper III** demonstrate that other materials from skin and space layer renovations such as steel, gypsum and plaster should be emphasized as well in future implementation of limit values in Sweden.

On a macro level, the results from this work highlight that there is a need for more comprehensive data collection from government agencies to support modeling of the built environment. The lack of complete bottom-up data on buildings and roads can be remedied using ML predictions, but ML models inevitably introduce uncertainties and errors in the prediction process. Therefore, a key takeaway message from this work is that data on physical dimensions of the built environment such as building footprint, usable floor space and height should be collected and made available by government agencies for research purposes and other analysis.

6 Future work

This section concludes this summary by summarizing the main findings of this work and providing an outlook for the future work.

6.1 Future work

As previously stated, the overall aim of this thesis is to improve estimations of material flows and embodied emissions from the Swedish built environment with a goal of decarbonization. However, the scope of this study is currently limited to roads and residential buildings, omitting several other components of the built environment. Therefore, the scope can be expanded in future studies to include more components of the built environment [7]. For transport infrastructures, the scope can be expanded to include railways, sidewalks, cycling paths, bridges and tunnels, and subways and tram ways. Furthermore, other infrastructures such as energy infrastructures (e.g., power grid and power generation technology) and supply and disposal infrastructures (e.g., pipelines and cables) could be included as well. For power generation technologies, Savvidou et al. [65] and Savvidou et al. (in preparation) already estimate material flows and embodied emissions from the construction and maintenance activities from wind turbines and PV panels in Sweden. These works can be expanded to cover all parts of the energy infrastructures. The main challenge in expanding the scope is to obtain inventory as well as MI data for the abovementioned infrastructures.

Furthermore, the exclusion of non-residential buildings represents a significant gap in the scope of this study. Non-residential buildings remain a significant challenge to model due to the lack of bottom-up inventory data. In addition, non-residential buildings are more heterogenous when it comes to their design and construction that require more material intensity data collection. The lack of inventory data could be tackled using the ML models developed in this work, but the limiting factor is that to the author's knowledge, there is currently no training dataset available for non-residential buildings. The building inventory data set used in this work does contain non-residential buildings, but none of the non-residential building entries contain usable floor space data. The ML models developed in this work require usable floor space for training and thus would not be possible to extend to non-residential buildings directly barring the release of new data.

Another approach to tackle the non-residential building inventory challenge could be using the recently published open-access building height datasets derived from remote sensing and ML

methods such as for example World Settlement Footprint 3D [66] and 3D-GloBFP [67]. A more systematic review of the literature on building height prediction can be carried out to determine the viability of using these datasets for the Swedish non-residential buildings. By incorporating non-residential buildings, a much clearer picture of the total embodied emissions from the Swedish construction sector can be derived.

The aim of this work is partially derived from the fact that there is currently no estimation of embodied emissions from the entire construction sector and its supply chain in Sweden. While detailed bottom-up modeling could provide more accurate estimates, such modeling is also time and resources intensive. Therefore, top-down approaches such as input-output analysis or inflow-driven DMFA could be used to model the remaining parts of the built environment. These top-down models can then be combined with the existing bottom-up models to estimate embodied emissions from the entire Swedish built environment. The combined model can thus be used to test emission reduction scenarios and inform policy making on the national level.

An alternative future work could be improving the model from **Paper III** to include more scenarios of future demand for housing. In the current study, only a business-as-usual scenario is investigated. One potential direction is to investigate sufficiency policies or scenarios that include measures such as increasing sharing of space, conversion of offices into residential buildings or from a more theoretical perspective of how a reduction in floor space per capita could reduce future embodied emissions. Sufficiency as a concept has recently received increased attention as supply-side emission reduction measures are not enough to reach net zero emissions as shown in this work. A potential direction to investigate sufficiency of the built environment stock while utilizing the results from this work is to adapt the consumption model developed by Pauliuk 2024 [68]. The consumption model uses the Lorenz curve concept to calculate levels of acceptable consumption while allowing some degree of overconsumption. As stated in the discussion of Pauliuk 2024 (Section 4.4), spatially explicit floor space data such as the one from **Paper II** could be matched with high resolution census data on grid cells to be ranked for Lorenz curves. The results from such work could then be used to inform future new construction scenarios and assess the possibilities for demand-side measures to reduce embodied emissions.

Another demand-side measure that has not been fully explored in this work is the circularity potential. A potential future work direction is to utilize the spatially explicit material stock and flow model to assess the circularity potential for each municipality in Sweden. A spatially

explicit analysis could highlight potential mismatches between inflows and outflows or inform locations of potential circularity hubs.

Finally, the building stock model can be further expanded and soft-linked with other models such as an agent-based model to analyze behaviors such as renovation decisions or potential for sufficiency measures. By incorporating behavioral factors into the model, the model and the results can be one step closer to reality. Regardless of the direction of future work, the overall aim of the work remains unchanged: to analyze and understand how to reduce the embodied carbon emissions from the Swedish built environment.

References

[1] UNEP, "2023 Global Status Report for Buildings and Construction: Beyond foundations - Mainstreaming sustainable solutions to cut emissions from the buildings sector," 2024. doi: 10.59117/20.500.11822/45095.

[2] OECD, *Global Material Resources Outlook to 2060: Economic Drivers and Environmental Consequences*. OECD, 2019. [Online]. Available: https://www.oecd-ilibrary.org/environment/global-material-resources-outlook-to-2060_9789264307452-en

[3] C. Camarasa *et al.*, "A global comparison of building decarbonization scenarios by 2050 towards 1.5–2 °C targets," *Nat. Commun.*, vol. 13, no. 1, pp. 1–11, 2022, doi: 10.1038/s41467-022-29890-5.

[4] H. L. Lou and S. H. Hsieh, "Towards Zero: A Review on Strategies in Achieving Net-Zero-Energy and Net-Zero-Carbon Buildings," *Sustain.* , vol. 16, no. 11, 2024, doi: 10.3390/su16114735.

[5] M. Röck *et al.*, "Embodied GHG emissions of buildings – The hidden challenge for effective climate change mitigation," *Appl. Energy*, vol. 258, p. 114107, 2020, doi: 10.1016/j.apenergy.2019.114107.

[6] D. Wiedenhofer *et al.*, "Mapping and modelling global mobility infrastructure stocks, material flows and their embodied greenhouse gas emissions," *J. Clean. Prod.*, vol. 434, no. February 2023, 2024, doi: 10.1016/j.jclepro.2023.139742.

[7] M. Lanau *et al.*, "Taking Stock of Built Environment Stock Studies: Progress and Prospects," *Environ. Sci. Technol.*, vol. 53, no. 15, pp. 8499–8515, 2019, doi: 10.1021/acs.est.8b06652.

[8] Q. Li, S. R. B. Gummidi, M. Lanau, B. Yu, and G. Liu, "Spatiotemporally Explicit Mapping of Built Environment Stocks Reveals Two Centuries of Urban Development in a Fairytale City, Odense, Denmark," *Environ. Sci. Technol.*, vol. 56, no. 22, pp. 16369–16381, 2022, doi: 10.1021/acs.est.2c04781.

[9] J. H. Arehart, F. Pomponi, B. D'Amico, and W. V Srubar, "A New Estimate of Building Floor Space in North America," *Environ. Sci. Technol.*, vol. 55, no. 8, pp. 5161–5170, 2021, doi: 10.1021/acs.est.0c05081.

[10] C. Wang, M. Ferrando, F. Causone, X. Jin, X. Zhou, and X. Shi, "Data acquisition for urban building energy modeling: A review," *Build. Environ.*, vol. 217, no. April, p. 109056, 2022, doi: 10.1016/j.buildenv.2022.109056.

[11] Q. Liu, J. Rootzén, and F. Johnsson, "Development of a machine learning model to improve

estimates of material stock and embodied emissions of roads," *Clean. Environ. Syst.*, vol. 14, no. June, 2024, doi: 10.1016/j.cesys.2024.100211.

[12] K. H. Rankin and S. Saxe, "A Future Growth Model for Building More Housing and Infrastructure with Less Embodied Greenhouse Gas," *Environ. Sci. Technol.*, vol. 58, no. 25, pp. 10979–10990, 2024, doi: 10.1021/acs.est.4c02070.

[13] I. Karlsson, J. Rootzén, A. Toktarova, M. Odenberger, F. Johnsson, and L. Göransson, "Roadmap for Decarbonization of the Building and Construction Industry—A Supply Chain Analysis Including Primary Production of Steel and Cement," p. 40, 2020.

[14] C. Fu, Y. Zhang, T. Deng, and I. Daigo, "The Evolution of Material Stock Research: From Exploring to Rising to Hot Studies," *J. Ind. Ecol.*, p. jiec.13195, 2021, doi: 10.1111/jiec.13195.

[15] D. B. Müller *et al.*, "Carbon emissions of infrastructure development," *Environ. Sci. Technol.*, vol. 47, no. 20, pp. 11739–11746, 2013, doi: 10.1021/es402618m.

[16] L. S. A. Rousseau, B. Kloostra, H. AzariJafari, S. Saxe, J. Gregory, and E. G. Hertwich, "Material Stock and Embodied Greenhouse Gas Emissions of Global and Urban Road Pavement," *Environ. Sci. Technol.*, vol. 56, no. 24, pp. 18050–18059, 2022, doi: 10.1021/acs.est.2c05255.

[17] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic, "A review of bottom-up building stock models for energy consumption in the residential sector," *Build. Environ.*, vol. 45, no. 7, pp. 1683–1697, 2010, doi: 10.1016/j.buildenv.2010.01.021.

[18] H. Tanikawa, T. Fishman, K. Okuoka, and K. Sugimoto, "The Weight of Society Over Time and Space: A Comprehensive Account of the Construction Material Stock of Japan, 1945-2010: The Construction Material Stock of Japan," *J. Ind. Ecol.*, vol. 19, no. 5, pp. 778–791, 2015, doi: 10.1111/jiec.12284.

[19] E. Müller, L. M. Hilty, R. Widmer, M. Schluep, and M. Faulstich, "Modeling Metal Stocks and Flows: A Review of Dynamic Material Flow Analysis Methods," *Environ. Sci. Technol.*, vol. 48, no. 4, pp. 2102–2113, 2014, doi: 10.1021/es403506a.

[20] D. Kong, A. Cheshmehzangi, Z. Zhang, S. P. Ardakani, and T. Gu, *Urban building energy modeling (UBEM): a systematic review of challenges and opportunities*, vol. 16, no. 6. Springer Netherlands, 2023. doi: 10.1007/s12053-023-10147-z.

[21] N. Milojevic-Dupon *et al.*, "OPEN EUBUCCO v0.1: European building Data Descriptor stock

characteristics in a common and open database for 200+ million individual buildings," 2023.

[22] European Environment Agency, "CORDA," 2023. https://corda.eea.europa.eu/_layouts/15/CustomLoginPageFBA/CustomLogin.aspx?ReturnUrl=%2F_layouts%2F15%2FAuthenticate.aspx%3FSource%3D%252Fsitepages%252Fhome%252Easpx&Source=%2Fsitepages%2Fhome.aspx

[23] European Comission, "EU Building Stock Observatory," 2023. https://ec.europa.eu/energy/eu-buildings-database_en

[24] Microsoft, "GlobalMLBuildingFootprints," 2023. https://github.com/microsoft/GlobalMLBuildingFootprints

[25] N. Milojevic-Dupont and F. Creutzig, "Machine learning for geographically differentiated climate change mitigation in urban areas," *Sustain. Cities Soc.*, vol. 64, no. October 2020, p. 102526, 2021, doi: 10.1016/j.scs.2020.102526.

[26] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[27] N. Milojevic-Dupont *et al.*, "Learning from urban form to predict building heights," *PLoS One*, vol. 15, no. 12, p. e0242010, 2020, doi: 10.1371/journal.pone.0242010.

[28] F. Nachtigall, N. Milojevic-Dupont, F. Wagner, and F. Creutzig, "Predicting building age from urban form at large scale," *Comput. Environ. Urban Syst.*, vol. 105, no. May, p. 102010, 2023, doi: 10.1016/j.compenvurbsys.2023.102010.

[29] D. B. Müller, "Stock dynamics for forecasting material flows—Case study for housing in The Netherlands," *Ecol. Econ.*, vol. 59, no. 1, pp. 142–156, 2006, doi: 10.1016/j.ecolecon.2005.09.025.

[30] P. K. Ohms, L. H. Horup, S. R. B. Gummidi, M. Ryberg, A. Laurent, and G. Liu, "Temporally dynamic environmental impact assessment of a building stock: Coupling MFA and LCA," *Resour. Conserv. Recycl.*, vol. 202, no. March 2023, p. 107340, 2024, doi: 10.1016/j.resconrec.2023.107340.

[31] V. Göswein, J. D. Silvestre, C. Sousa Monteiro, G. Habert, F. Freire, and F. Pittau, "Influence of material choice, renovation rate, and electricity grid to achieve a Paris Agreement-compatible building stock: A Portuguese case study," *Build. Environ.*, vol. 195, p. 107773, 2021, doi: 10.1016/j.buildenv.2021.107773.

[32] P. Berrill and E. G. Hertwich, "Material Flows and GHG Emissions from Housing Stock Evolution in US Counties, 2020–60," *Build. Cities*, vol. 2, no. 1, pp. 599–617, 2021, doi:

10.5334/bc.126.

[33] P. Berrill, E. J. H. Wilson, J. L. Reyna, A. D. Fontanini, and E. G. Hertwich, “Decarbonization pathways for the residential sector in the United States,” *Nat. Clim. Chang.*, vol. 12, no. 8, pp. 712–718, 2022, doi: 10.1038/s41558-022-01429-y.

[34] S. Brand, *How buildings learn: What happens after they’re built*. Penguin, 1995.

[35] J. Persson, “Sweden’s long-term strategy for reducing greenhouse gas emissions,” p. 87, 2020.

[36] Trafikverket, “Requirements for Reducing Greenhouse Gas Emissons,” p. 2.

[37] Ministry of Infrastructure, “Sweden’s Third National Strategy for Energy Efficient Renovation ,” no. May, pp. 5–78, 2020, [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiis9iuu5r8AhWCQvEDHXfjDOcQFnoECA4QAQ&url=https%3A%2F%2Fenergy.ec.europa.eu%2Fdocument%2Fdownload%2Facc3e1f5-f0ef-457c-a4e7-f41ec7198deb_en%3Ffilename%3Dse_2020_ltrs_o

[38] Trafikverket, “Lastkajen – Sveriges väg- och järnvägsdata,” *Trafikverket*, 2022. https://www.trafikverket.se/tjanster/data-kartor-och-geodatatjanster/hamta-var-oppna-data/lastkajen---sveriges-vag--och-jarnvagsdata/

[39] N. Japkowicz, “Learning from imbalanced data sets: a comparison of various strategies,” *AAAI Work. Learn. from Imbalanced Data Sets*, pp. 0–5, 2000.

[40] Boverket, “Miljonprogrammet,” 2024. https://www.boverket.se/sv/samhallsplanering/stadsutveckling/miljonprogrammet/

[41] G. Dong and H. Liu, *Feature engineering for machine learning and data analytics*. CRC press, 2018.

[42] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. “ O’Reilly Media, Inc.,” 2018.

[43] M. Berghauser Pont *et al.*, “The spatial distribution and frequency of street, plot and building types across five European cities,” *Environ. Plan. B Urban Anal. City Sci.*, vol. 46, no. 7, pp. 1226–1242, 2019, doi: 10.1177/2399808319857450.

[44] J. F. Rosser, D. S. Boyd, G. Long, S. Zakhary, Y. Mao, and D. Robinson, “Predicting residential building age from map data,” *Comput. Environ. Urban Syst.*, vol. 73, pp. 56–67, 2019, doi: 10.1016/j.compenvurbsys.2018.08.004.

[45] M. Fleischmann, “momepy: Urban Morphology Measuring Toolkit,” *J. Open Source Softw.*,

vol. 4, no. 43, p. 1807, 2019, doi: 10.21105/joss.01807.

[46] P. Zhang, Y. Jia, and Y. Shang, “Research and application of XGBoost in imbalanced data,” *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, 2022, doi: 10.1177/15501329221106935.

[47] G. LemaÃŽtre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[48] I. Mani and I. Zhang, “kNN approach to unbalanced data distributions: a case study involving information extraction,” in *Proceedings of workshop on learning from imbalanced datasets*, ICML, 2003, pp. 1–7.

[49] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” in *Icml*, Citeseer, 1997, p. 179.

[50] J. Laurikkala, “Improving identification of difficult small classes by balancing class distribution,” in *Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8*, Springer, 2001, pp. 63–66.

[51] N. Milojevic-Dupont *et al.*, “Learning from urban form to predict building heights,” *PLoS One*, vol. 15, no. 12 December, pp. 1–22, 2020, doi: 10.1371/journal.pone.0242010.

[52] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[53] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[54] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[55] G. Ke *et al.*, “LightGBM: A highly efficient gradient boosting decision tree,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[56] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “Catboost: Unbiased boosting with categorical features,” *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.

[57] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2623–2631, 2019, doi: 10.1145/3292500.3330701.

[58] M. Lanau and G. Liu, “Developing an Urban Resource Cadaster for Circular Economy: A Case of Odense, Denmark,” *Environ. Sci. Technol.*, vol. 54, no. 7, pp. 4675–4685, 2020, doi: 10.1021/acs.est.9b07749.

[59] S. Pauliuk and N. Heeren, “ODYM—An open software framework for studying dynamic material systems: Principles, implementation, and data structures,” *J. Ind. Ecol.*, vol. 24, no. 3, pp. 446–458, 2020, doi: 10.1111/jiec.12952.

[60] I. Sartori, N. H. Sandberg, and H. Brattebø, “Dynamic building stock modelling: General algorithm and exemplification for Norway,” *Energy Build.*, vol. 132, pp. 13–25, 2016, doi: 10.1016/j.enbuild.2016.05.098.

[61] X. Zhong *et al.*, “Global greenhouse gas emissions from residential and commercial building materials and mitigation strategies to 2060,” *Nat. Commun.*, vol. 12, no. 1, p. 6126, 2021, doi: 10.1038/s41467-021-26212-z.

[62] S. Deetman, S. Marinova, E. van der Voet, D. P. van Vuuren, O. Edelenbosch, and R. Heijungs, “Modelling global material stocks and flows for residential and service sector buildings towards 2050,” *J. Clean. Prod.*, vol. 245, p. 118658, 2020, doi: 10.1016/j.jclepro.2019.118658.

[63] L. Niamir, A. Mastrucci, and B. van Ruijven, “Energizing building renovation: Unraveling the dynamic interplay of building stock evolution, individual behaviour, and social norms,” *Energy Res. Soc. Sci.*, vol. 110, no. July 2023, p. 103445, 2024, doi: 10.1016/j.erss.2024.103445.

[64] Boverket., *Limit values for climate impact from buildings and an expanded climate declaration.* 2023. [Online]. Available: https://www.boverket.se/globalassets/engelska/limit-values-for-climate-impact-from-buildings-and-an-expanded-climate-declaration.pdf

[65] G. Savvidou and F. Johnsson, “Material Requirements, Circularity Potential and Embodied Emissions Associated with Wind Energy,” *Sustain. Prod. Consum.*, vol. 40, no. July, pp. 471–487, 2023, doi: 10.1016/j.spc.2023.07.012.

[66] T. Esch *et al.*, “World Settlement Footprint 3D - A first three-dimensional survey of the global building stock,” *Remote Sens. Environ.*, vol. 270, no. January, p. 112877, 2022, doi: 10.1016/j.rse.2021.112877.

[67] Y. Che *et al.*, “3D-GloBFP : the first global three-dimensional building footprint dataset,” *Earth Syst. Sci. Data*, vol. 11319912, no. June, pp. 1–28, 2024.

[68] S. Pauliuk, “Decent living standards, prosperity, and excessive consumption in the Lorenz

curve," *Ecol. Econ.*, vol. 220, no. January, p. 108161, 2024, doi: 10.1016/j.ecolecon.2024.108161.