Development of a machine learning model to improve estimates of material stock and embodied emissions of roads

Cleaner Environmental Systems 14 (2024) 100211

Contents lists available at ScienceDirect

Cleaner Environmental Systems

journal homepage: www.journals.elsevier.com/cleaner-environmental-systems

Development of a machine learning model to improve estimates of material stock and embodied emissions of roads

Qiyu Liu [a,*], Johan Rootzén [b], Filip Johnsson [a]

[a] Department of Space, Earth and Environment, Chalmers University of Technology, 421 96, Gothenburg, Sweden
[b] IVL Swedish Environmental Research Institute, Aschebergsgatan 44, 411 33, Gothenburg, Sweden

A B S T R A C T

Material flow analysis is an important tool for estimating material flows and embedded emissions of transport infrastructure. Missing attributes tend to be a major barrier to accurate estimates. In this study a machine learning model is developed to estimate the missing data in a statistics dataset of roads, to enable a bottom-up material stock and flow analysis. The proposed approach was applied to the Swedish road network to predict missing data for road width in the statistical dataset. The predicted hybrid dataset was then used to estimate material stocks, flows, and embodied emissions from Year 2020 to Year 2045 using decarbonization scenarios with a supply chain perspective. The study demonstrates that machine learning models can be used to enable national-level material stock and flow analyses of roads. Multiple machine learning algorithms were tested, and the best performing model achieved an R^2 value of 0.784. In the scenario-based analysis, the embodied emissions of Swedish roads could be reduced by up to 51% using available materials.

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) 6th assessment report concludes that existing Nationally Determined Contributions (NDCs) are likely to result in global warming far exceeding 1.5 °C, and that limiting warming to less than 2 °C will require rapid intensification of mitigation efforts after Year 2030 (IPCC, 2022). The global construction sector has a significant role to play in societal decarbonization, given that the sector is responsible for 36% of total energy consumption and 39% of the carbon dioxide (CO_2) emissions related to energy and industrial processes (UN Environment Programme, 2019). Echoing the IPCC report, it has been suggested that the construction sector needs to become carbon-neutral or even have negative emissions after Year 2030 (Rockström et al., 2017), which represents a formidable challenge.

Infrastructure construction, including road construction, accounts for a significant share of the carbon footprint of the global construction sector. Müller et al. (2013) in a study from 2013 estimated the carbon footprint of the existing global infrastructure stock in Year (2008) as 122 (−20/+ 15) $GtCO_2$. More recently, Rousseau et al. (2022) estimated embodied greenhouse gas (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). 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 (Wenz et al., 2020). 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 (Nasir et al., 2021).

Material stock and flow analysis (MFA) is a well-developed methodology to estimate the stock and flow of construction materials (for a review, see (Augiseau and Barles, 2017a)). MFA studies can be classified as using either a top-down or bottom-up approach (Augiseau and Barles, 2017b). Ebrahimi et al. (2022) have conducted a literature review on material flow analysis (MFA) studies of transport infrastructure. The review indicates that most of the recent studies on transport infrastructures have used the bottom-up approach. Compared to a top-down approach, the bottom-up approach helps to identify the composition of the stock in a more-detailed manner (Tanikawa et al., 2015). The disadvantage of the bottom-up approach is that it requires more data and labor, as each item in the inventory needs to be quantified (Lanau et al., 2019).

The lack of data is a major challenge for the expansion of both top-down and bottom-up MFA studies of transport infrastructures such as roads (Lanau et al., 2019) (Nguyen et al., 2019). At the global level, the Global Roads Inventory Project (GRIP) has gathered and harmonized information related to road length and type of roads for 222 countries (Meijer et al., 2018). Rousseau et al. (2022) 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

* Corresponding author.
E-mail address: qiyu@chalmers.se (Q. Liu).

https://doi.org/10.1016/j.cesys.2024.100211
Received 7 March 2024; Received in revised form 19 June 2024; Accepted 11 July 2024
Available online 14 July 2024

have the lowest levels of missing data. Similarly, OpenStreetMap, which is another global-level dataset describing roads, suffers from data incompleteness (Funke et al., 2015). In Europe, many gaps exist in the official Eurostat statistics on transport infrastructure (Eurostat, 2022). In addition, non-government-owned roads, such as communal roads, have overall lower data quality and are sometimes not included in the national statistics (Wiedenhofer et al., 2015).

To overcome this lack of data, a variety of approaches have been adopted in MFA studies of transport infrastructures. To limit the challenges associated with low data quality and data unavailability, the scope of the study can be narrowed, as has been done in several works (Tanikawa and Hashimoto, 2009; Guo et al., 2014, 2017; Currie et al., 2017; Gontia et al., 2019; Lanau and Liu, 2020; Miatto et al., 2021; Kloostra et al., 2022; Khumvongsa et al., 2023). The most-common methods for imputing missing data are interpolation (Wiedenhofer et al., 2015), (Miatto et al., 2017), (Deng et al., 2022) and extrapolation (Han and Xiang, 2013), (Wiedenhofer et al., 2015). The main drawback of these methods is that they do not capture the inherent heterogeneity of the physical properties of roads, since they use extensive data aggregation (Wang et al., 2022). This motivates the development of novel methods that can utilize a wider range of available data, to complement and improve the predictions of missing data with finer geographic scope.

Machine learning is an emerging method to estimate the stocks and flows of materials across various geographic scales (Donati et al., 2022). Several recent studies have applied machine learning-based approaches to estimate the material stocks and flows of roads by predicting various types of road attributes (Ebrahimi et al., 2022), (Bao et al., 2023), (Zhang et al., 2023). Zhang et al. (2023) employed a set of time series analysis-based machine learning models to project the historical material stock (MS) of Japanese roads from Year (2020) to Year 2050 under five different national shared socio-economic pathways (SSPs). The study used an archetype-based approach by dividing roads into archetypes and applying a material intensity (MI) to estimate the material stocks and flows. The road data were collected for each prefecture in Japan, including national, prefectural, and municipal roads. The explanatory variables used were gross domestic product (GDP), population growth, and transportation statistics. The main limitation of this approach is that the MS is projected at a too-aggregated level of geographic resolution (prefectures). The archetype-based approaches are also unable to capture fully the effects of traffic and climate on the stocks and flows of roads (Wang et al., 2022).

Another strand of research aims to overcome these limitations by using machine learning models to predict the depth of the road layers. Ebrahimi et al. (2022) estimated and predicted the material stocks and flows of the Norwegian road network by predicting the depth of roads with a decision tree-based machine learning 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 machine learning training process requires extensive data, the analyses were limited to national roads, for which all the input data were complete Similarly, Wang et al. (2022) estimated the material stocks and flows of road infrastructures in Belgium using a combination of machine learning models and the archetype-based approach. The missing layer thicknesses of asphalt motorways were predicted using a machine learning approach, while the layer thicknesses of other road types were predicted using an archetype approach. The reason for using this hybrid approach is that roads that are not asphalt highways lack the necessary data to train the machine learning model.

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. (2022)), and this impedes the implementation of bottom-up MFA 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 (Wiedenhofer et al., 2015). A key data-point for estimating the material stock and flows of roads is the road width, since it has been demonstrated that the material stock of roads is highly sensitive and varies significantly with road width (Yu et al., 2021).

By developing and applying a machine learning model that can predict missing road width data this study contributes to ongoing method development aimed at improving the accuracy in MFA studies to estimate the material stock and embodied emissions of road infrastructure construction. The novelty of this study is therefore to explore and show how machine learning methods can fill in the gaps in 2D road data for non-government-owned roads using existing open administrative data for government owned roads at a high spatial resolution. Thus, we address the aforementioned gap in literature where non-government-owned roads cannot be represented with reasonable accuracy in road MFA studies (Wiedenhofer et al., 2015). Additionally, we utilize street-network features which have not been used in previous studies. Furthermore, this study proposes a framework for how to address the issue of lack of data for the purpose of estimating embodied carbon emissions of roads using bottom-up MFA models. The model is demonstrated for the Swedish road stock but can be applied to other geographical areas facing similar problems of incomplete attribute data. The proposed framework may be used for analyzing pathways and policies to reduce embodied carbon and thus should be of interest to policy makers and local stakeholders.

The present study aimed to develop a novel machine-learning based method to predict missing road width data, for the purpose of estimating the material stock of the Swedish road system and using this as the basis for the calculation of embodied carbon emissions. The developed machine learning model uses roads that have no missing data as the training dataset and testing dataset. The resulting machine learning model is then used to make predictions for the roads that have missing data. The completed dataset is used to conduct a stock-driven MFA up to the Year 2045. Scenario-based emission factors are applied to the inflows to estimate the embodied emissions and their potential reduction pathways.

The paper is structured as follows. In Section 2, we describe the scope and method of this study. In Section 3, we present our findings. Section 4 discusses the applicability of machine learning methods to MFA studies for roads and the limitations of the study. Finally, in Section 5, we draw conclusions and outline key areas for further investigation and for policy actions.

2. Materials and methods

This study proposes a new machine learning-based method to estimate missing road width data in a statistical dataset, for the purpose of material stock and flow quantification. The proposed method is demonstrated for a Swedish national road dataset utilizing open-source data and software. This section presents the system boundary, data sources, methodologic framework, data preprocessing, feature engineering, machine learning models, and estimation of the material stock and flow.

2.1. Methodologic approach

The methodologic approach developed and applied in this work is divided into four steps: 1) Data gathering and preprocessing; 2) Feature engineering; 3) Machine learning; and 4) Material stock and flow analysis Fig. 1. The methodology is developed based on the assumption that the physical attributes of roads within a region are correlated to some extent. Transport infrastructures are constrained by the natural conditions of the location and the physical attributes of a road or other type of infrastructure that reciprocally influence the spatial characteristics of the location (Rodrigue, 2020). For example, a stretch of road that is surrounded by apartment buildings is more likely to be a paved road, whereas a stretch of road on farmland is more likely to be unpaved

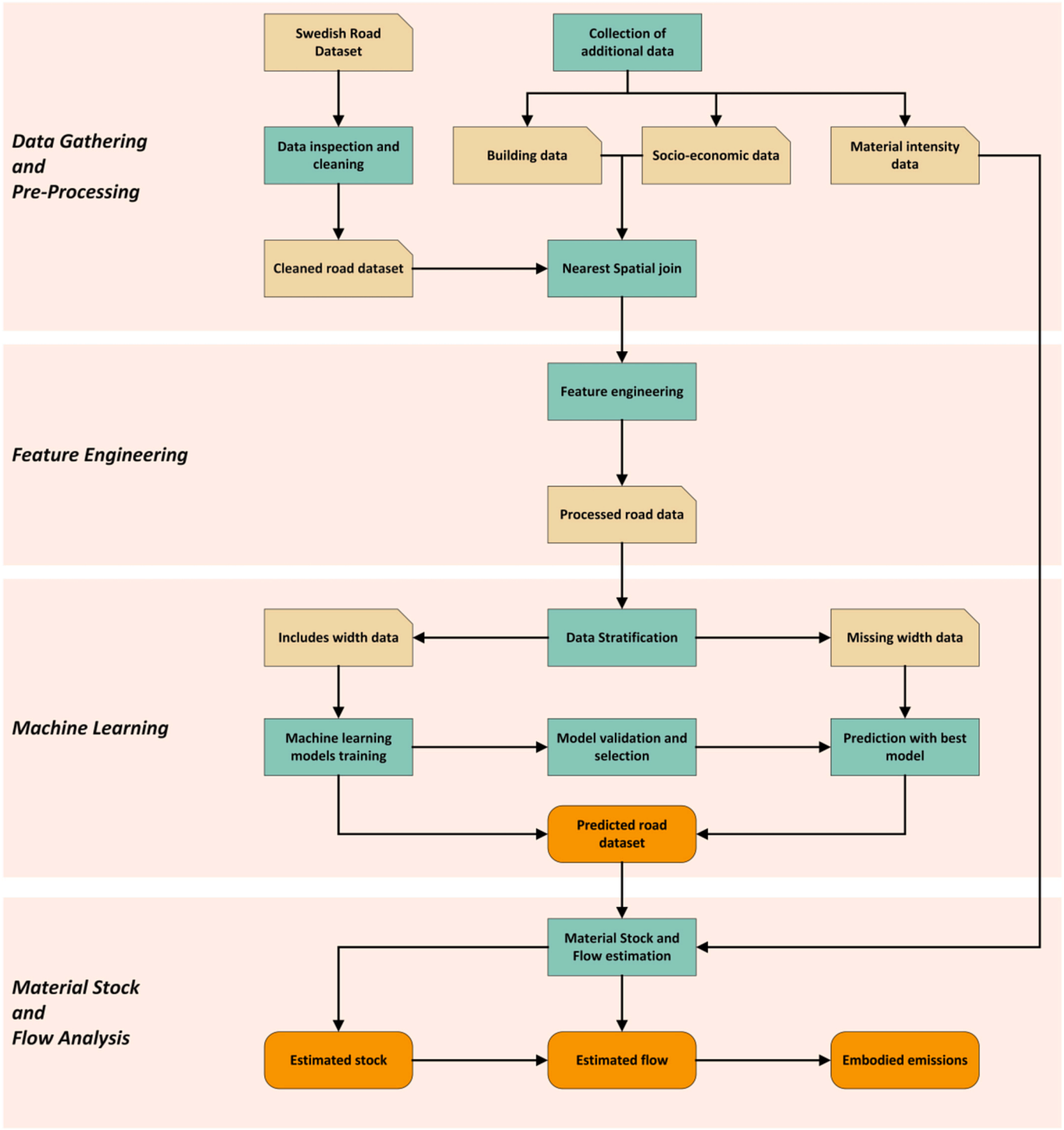

Fig. 1. Flow diagram describing the main steps in the analysis work: 1) Data gathering and data preprocessing; 2) Feature engineering; 3) Machine learning; and 4) Material stock and flow analysis. The light-brown boxes represent input data, the green boxes represent intermediate processing steps in the workflow, and the orange-colored boxes represent the outputs. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

or be a gravel road. In practice, this means that when using known spatial characteristics and physical attributes of roads in an incomplete dataset, the missing information can be estimated using a machine learning regression model.

2.2. Data gathering and pre-processing

This section describes the first step of the methodologic framework, as shown in Fig. 1.

2.2.1. Data gathering and data sources

This study used the following open-source datasets.

1) *Swedish road shapefile with attributes from the Swedish Transport Administration (STA)* (Trafikverket, 2022a). The road shapefile includes 1-dimensional line strings that represent the shapes of the road sections with coordinates, a unique ID, geometric length, geometric width, road owner type (state-owned, municipally owned or

privately owned), road type, road surface type, and the speed limit for each road section.

2) *Swedish building footprint shapefile from the Swedish Land Survey* (Lantmäteriet, 2023a), which contains 2-dimensional footprints of buildings in Sweden, the type of building, the perimeter, and the area of each footprint.
3) *Socio-economic data from Statistics Sweden* (Statistikmyndigheten, 2023a), which were used as additional predictor variables in the machine learning process. The socio-economic data includes administrative and population data. The administrative data include all counties, municipalities, demographic statistical areas (DeSO) (Statistikmyndigheten, 2023b), and regional statistical areas (RegSO) (Statistikmyndigheten, 2023c) in Sweden, in the form of GIS shapefiles. Sweden is administratively divided into 21 counties, 290 municipalities, 5984 demographic statistical areas, and 3363 regional statistical areas. The population data break down Sweden's population into 1-km grids, accessed in the form of a GIS shapefile (Statistikmyndigheten, 2023d).
4) *Material intensity (MI) data for road archetypes from the STA* (Trafikverket, 2022b)*, and road lifetime distributions of the roads in Sweden* (Nilsson et al., 2020). These were used for the MFA. The material intensity (MI) coefficient data (material use per m^2 of road) were obtained from the STA's Klimatkalkyl tool (Trafikverket, 2022b). The Klimatkalkyl tool has been designed for construction companies in Sweden that work with the STA to calculate the life-cycle energy use and climate impact of new construction projects. This database contains the MIs of the different archetypes of roads that were used to categorize the stock data. These MIs were obtained from previously executed construction projects. In some cases, the MIs were computed from the results obtained from multiple construction projects. The MI values are listed in Table 1.

Table 2 presents the lifetime distribution parameters of the roads. The lifetimes of roads are assumed to follow a different Weibull distribution for each region in Sweden. Svenson et al. (2016) have developed the lifetimes using a mixed proportional hazards model that applies independent variables, such as climate zone, bearing capacity class and speed limit. The lifetime distributions are assumed to be the same for all types of roads located in the same geographic region. The lifetime distributions are used to model maintenance of roads, and we assume that no existing roads are demolished.

2.2.2. Data preprocessing

In the data preprocessing phase, all the input data were verified and processed in Python, and all the GIS-related calculations were performed using the GeoPandas package (Jordahl et al., 2019). As described above, the Swedish Traffic Administration (STA) collects extensive data on state-owned roads, whereas the datasets for municipally owned and privately owned roads have incomplete attributes (Liljenströ et al., 2019). The road dataset contains 2,003,127 sections of roads, corresponding to approximately 1,114,879 km of paved and gravel roads. An assessment of the dataset showed that the majority of missing road width data are associated with privately owned roads, as shown in Fig. 2. For an overview of the data used in the analysis see Fig. 4.

Table 1
Material intensity values for each road archetype, based on data from the Klimatkalkyl tool of the Swedish Transport Administration (Trafikverket, 2022b).

	Asphalt (tonne/m^2)	Steel (tonne/m^2)	Gravel (tonne/m^2)
One-lane Road	0.403	0	0.499
Two-lane road	0.403	0	0.473
Highway	0.403	0.04	0.454
Meeting free road	0.403	0.0097	0.454
Gravel road	0	0	0.6

Table 2
Lifetime Weibull distribution parameters for each region in Sweden, based on data from Nilsson et al. (2020). The scale parameter is 3.0199 for all regions.

Regions	Shape
North	13.93
Middle	12.60
Stockholm	10.95
South	13.25
East	11.29
West	13.38

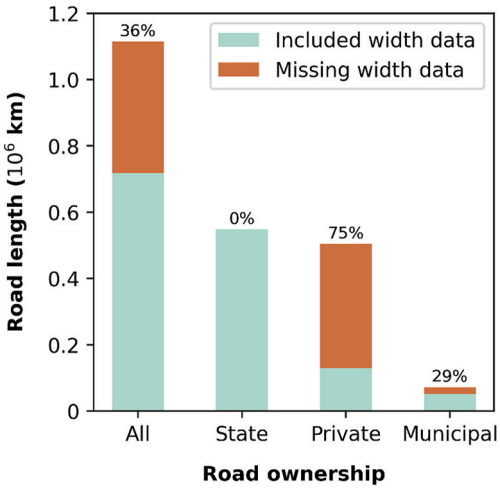

Fig. 2. Lengths and percentages of roads with missing road width data according to ownership type, based on data from the Swedish Land Survey (Lantmäteriet, 2023b).

The building footprints were joined with the road shapefile using the 'nearest spatial join' function in GeoPandas, with the distance between the matching road and building being computed and stored as an attribute. In addition, the geometries of the matching building were retained as attributes. The socio-economic data were similarly combined with the road shapefile using the 'nearest spatial join' function in GeoPandas, as additional attributes. Further details of data preprocessing steps can be found in the *Supplementary Information*.

2.3. Feature engineering

In the context of machine learning: "a *feature* is a numeric representation of an aspect of raw data, and *feature engineering* is the act of extracting features from raw data and transforming them into formats that are suitable for the machine learning model" (Zheng and Casari, 2018). For regression tasks, a feature is equivalent to a predictor variable. The aim of the feature engineering phase of this work was to generate useful and useable predictor variables so that the regression model could make better predictions on the response variables. In total, 29 features were used; the list of features and their descriptions are given in Table 3. The details of the features and the assumptions underlying their selection are explained in this section.

2.3.1. Road and building features

In the National Road Database (NVDB) database, the geometry of roads is represented as a 1-dimensional line string, which limits the possibilities for generating geometric features for a road dataset. A geometric feature is a numeric description of a given geometry, such as a perimeter or area of a 2-dimensional polygon, or the height of a 3-dimensional object. Three geometric features were computed based on

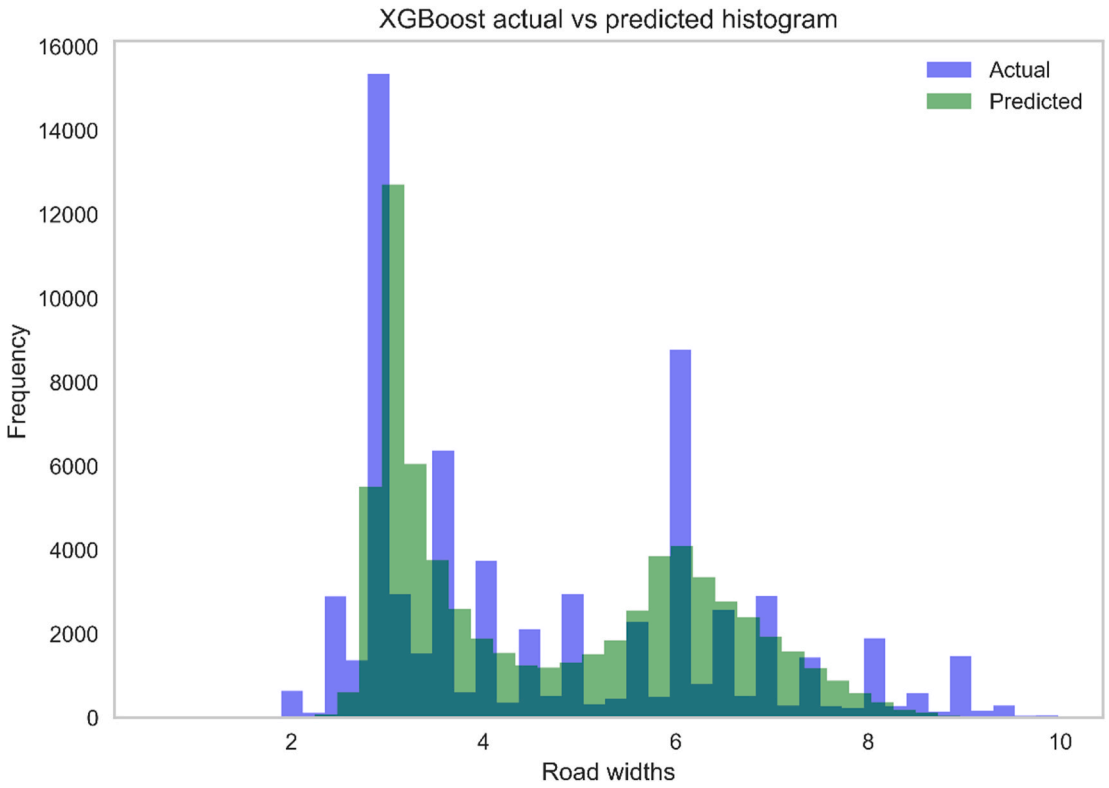

Fig. 3. Comparison of actual versus predicted value in meters, the number of bins used is 40.

the line strings: 1) convex hull area; 2) convex hull perimeter; and 3) envelope area. The relatively low number of features limited the explanatory power of the regression model. Therefore, additional data were collected and added to the road dataset.

The patterns and characteristics of a road network correlate with and are influenced by the spatial distribution of its surrounding buildings and land-use pattern (Kasraian et al., 2016). The assumption is made that the characteristics of the building nearest to the road correlate with the width of the road. Four nearest-building-based features were added, including: 1) distance to the nearest building; 2) nearest building perimeter; 3) nearest building area; and 4) nearest building type.

2.3.2. *Socio-economic features*

The assumption made regarding the addition of socio-economic features is that transport infrastructures in the same geographic area tend to correlate with each other (Rodrigue, 2020). As described above, these socio-economic features contain information as to the region or area in which a road section is located. The socio-economic features added can be divided into administrative and population features.

The four administrative features or areas are: 1) county; 2) municipality; 3) demographic statistical area (DeSO); and 4) regional statistical area (RegSO). Demographic statistical areas are subdivisions of municipalities that take geographic conditions into account. Regional statistical areas are subdivisions of municipalities that are used for statistical monitoring of socio-economic segregation. Population density and distribution have a direct correlation with the pattern of a road network (Zhao et al., 2016). Therefore, the addition of population features aims to capture this correlation.

2.3.3. *Network features*

Urban morphology refers to the study and analysis of the physical form, layout, structure, and evolution of an urban area or a city (Kropf, 2018). The topological and morphological characteristics of a road network can be analyzed computationally using a graphical approach (Jiang and Claramunt, 2004). In similarity to the socio-economic features, the assumption is that these characteristics have complex correlations with the geometric attributes of roads, which could be captured by a machine learning model.

These network features are computed using the Python package Momepy (Fleischmann, 2019), which is a flexible tool for computing urban morphometric characters. The list of features computed using Momepy are: 1) Local closeness centrality (400-m radius); 2) Linearity; 3) Neighboring Street orientation deviation; 4) Connectivity Gamma; 5) Edge node ratio; 6) Mean node degree; 7) Mean node distance; 8) Segments length; 9) Proportion of three-way intersection; 10) Proportion of four-way intersection; 11) Proportion of dead ends; and 12) Meshedness. All features are computed for each road line string.

2.4. *Machine learning*

The underlying correlation between the spatial structure of a road's location and its physical attributes is both complex and non-linear. Therefore, this study employed a supervised machine -learning regression approach, which uses the proportion of data with known width to train and test machine learning models, so as to capture the correlation; thereafter, the trained model is used to predict the missing width data, as shown in Fig. 2. Multiple machine learning models were trained, tested, and validated for this study, including Random Forest (RF), extreme gradient boosting (XGBoost), CatBoost, and Light gradient-boosting machine (LightGBM).

Random Forest is an often-used machine learning algorithm that combines the principles of decision trees and ensemble learning. A decision tree is a tree-like structure in which each node represents a feature, each edge represents a decision based on that feature, and each leaf node represents a predicted outcome. An RF consists of multiple decision trees, and each tree is built using a different subset of training data selected using the bootstrapping technique and a subset of randomly selected features. This introduces diversity to the tree and reduces overfitting. The RF model is implemented using the Python scikit-learn package (ver. 1.3.0) (Pedregosa et al., 2011).

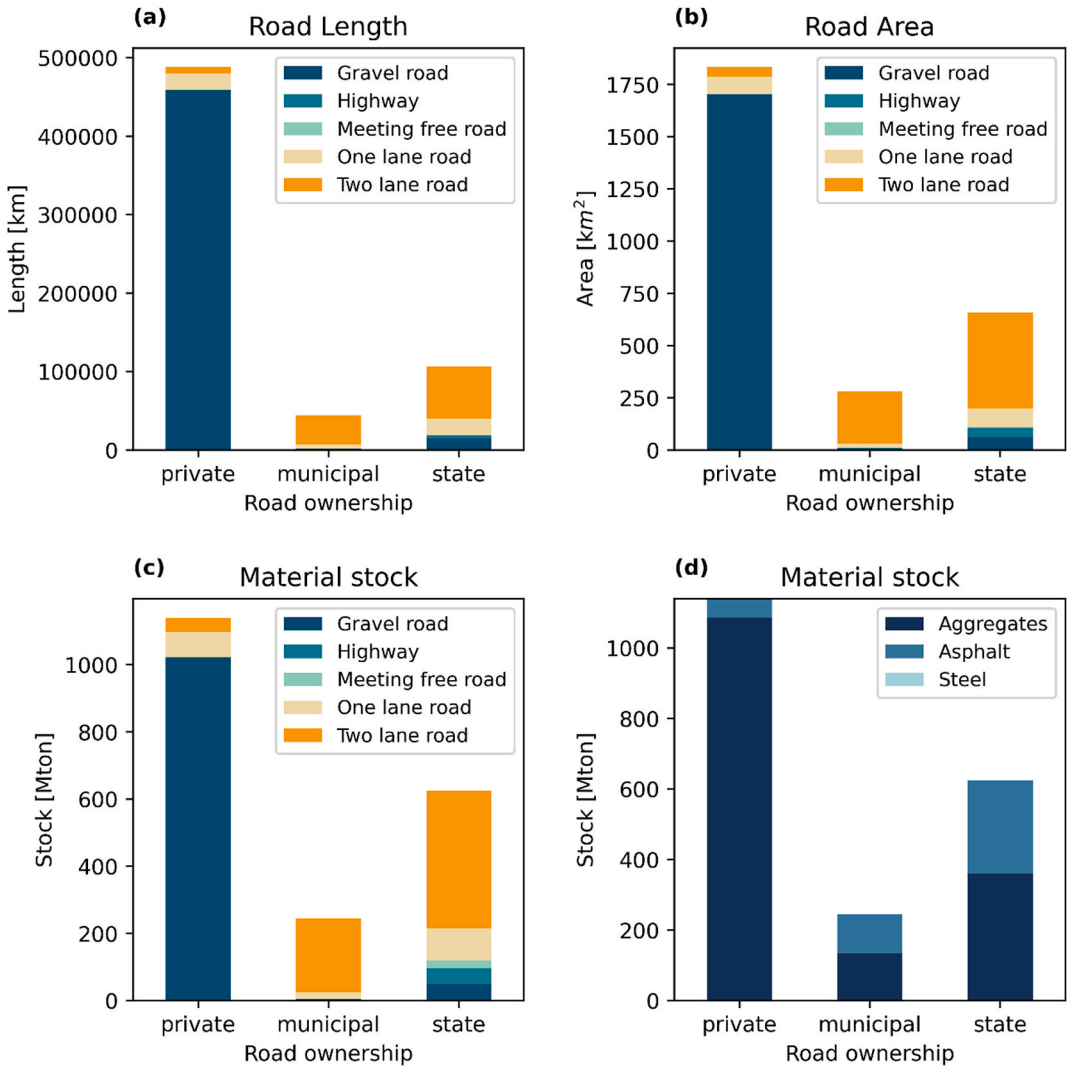

Fig. 4. Comparison of road sections based on the type of road and type of road ownership for the year 2020 in Sweden: a) road length (km) per road type; b) road area (km^2); c) material stock by road type (Mton); and d) material stock by material type (Mton).

XGBoost is a machine learning algorithm that is a member of the family of gradient-boosting techniques (Chen and Guestrin, 2016). 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.

CatBoost is another gradient-boosting algorithm that is known for its ability to handle categorical features without requiring extensive preprocessing (Prokhorenkova et al., 2018). CatBoost uses an ordered boosting technique, which helps to reduce overfitting by controlling the magnitude of individual trees' contributions to the ensemble. This is achieved by introducing an additional regularization term into the optimization process. A specialized algorithm is used to calculate the gradients for categorical features in the boosting process.

Similar to XGBoost and CatBoost, LightGBM is based on the gradient-boosting framework (Ke et al., 2017). It builds an ensemble of weak learners to create a strong predictive model. One of the key features of LightGBM is its leaf-wise tree growth strategy. Unlike the depth-wise growth used in other boosting algorithms, LightGBM grows trees in a leaf-wise fashion. This means that the algorithm chooses the leaf with the largest gradient for expansion at each step. This approach can lead to faster convergence and potentially more-accurate models. LightGBM also uses histogram-based learning for splitting nodes during tree construction. This technique discretizes the feature values into bins, which reduces memory usage and speeds up the process of finding the best split.

During the training and testing processes, the proportion of data with known width data is divided into response variable (width) and predictor variables (all other attributes) as inputs to the regression model. The response and predictor variables are further split into training and testing datasets with a 80/20 split using the Python scikit-learn package (ver. 1.3.0) (Pedregosa et al., 2011). The performance evaluation metrics used during the model validation process are mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). Furthermore, the computational time for each model is recorded for model selection. Lastly, the hyperparameters of each model are tuned to improve model performance using the Python Optuna package (ver. 3.3.0) (Akiba et al., 2019). A 5-fold cross-validation procedure with shuffle is employed for each model during the hyperparameter tuning process, to minimize overfitting.

The tuned and cross-validated results for each model are compared, and the model with the best overall performance is selected. To create a new hybrid dataset consisting of real and synthetic road widths, the best model is used to predict the missing width values, and the predicted width values are then appended to the proportion of data with width values.

Table 3
Description of the features selected for further use in the machine learning step.

Features	Description
Road features	
Road ownership	The ownership of the road section (state, municipality or private)
Road surface type	The surface type of the road section (asphalt paved or gravel)
Road archetype (m^2)	The type of the road section (e.g., highway, one-lane road, or two-lane road)
Road length (m)	The length of each road section
Speed limit (km/h)	The upper speed limit of the road section
Convex hull area (m^2)	The area of the smallest convex polygon containing all points in the line string
Convex hull perimeter (m)	The perimeter of the smallest convex polygon containing all points in the line string
Envelope area (m^2)	The area of the envelope of the road section line string
Building features	
Distance to the nearest building (m)	The distance between the center point of the road (line string) and the center point of the nearest building (polygon), computed using GeoPandas
Nearest building perimeter (m)	The perimeter of the nearest building to the road section
Nearest building area (m^2)	The area of the nearest building to the road section
Nearest building type	The type of the nearest building to the road section (e.g., detached building, apartment building or office)
Socio-economic features	
County	The county in which the road section is located
Municipality	The municipality in which the road section is located
Demographic statistical area (DeSO)	The demographic statistical area in which the road section is located
Regional statistical area (RegSO)	The regional statistical area in which the road section is located
Population	The number of people living inside the 1 km^2 grid in which the road section is located
Network features	
Local closeness centrality (400 m radius)	The average distance to every other road intersection (node) from each intersection in a road network (graph)
Linearity	The Euclidean distance of the road section divided by the road length
Neighboring street orientation deviation	The mean deviation of solar orientation of adjacent roads
Gamma	The connectivity gamma index for the road network (graph) around each road intersection (node)
Edge node ratio	The ratio of edges and intersections (node) for the intersection (node)
Mean node degree	The mean node degree around each intersection (node) for the whole road network (graph)
Mean node distance	The mean distance to neighboring intersections (node)
Segments length	The mean length of each segment
Proportion of three-way intersection	The proportion of three-way intersections in the road network (graph)
Proportion of four-way intersection	The proportion of four-way intersections in the road network (graph)
Proportion of dead ends	The proportion of dead ends in the road network (graph)
Meshedness	The meshedness for the road network (graph) around each intersection (node)

2.5. Sweden as a case study

Sweden is located in Northern Europe with a total land area of 447,430 km^2 and a total population of 10.5 million[1] (Statistikmyndigheten, 2023a). Sweden has a relatively low population density and one of the greatest road lengths per capita in the EU (Brons et al., 2022). At the national level, Sweden has committed to reduce national GHG emissions to net-zero by Year (2045) (Persson, 2020), which is why the scenario analysis in this study is up to Year 2045. The STA has announced its ambition for Sweden's transport infrastructure to be carbon-neutral by Year (2040) (Trafikverket). All roads in Sweden are recorded in NVDB, with varying degrees of completeness depending on road ownership. To achieve the ambitious emissions reduction goal, it is necessary to understand what the future material flows might look like. This makes Sweden an ideal case study to investigate how to overcome shortcomings in the available dataset for MFA and to demonstrate its application with regards to embodied emissions reductions.

[1] Population size as of February 2023.

2.6. System boundary

The study includes all paved and gravel roads in Sweden (as of Year, 2023) that are owned by three types of actors: the STA, local municipal governments, and private owners or owner associations. This study focuses on the road network and excludes sidewalks, cycleways, roundabouts (traffic circles), tunnels, and bridges. It covers all layers of the road, except for the foundations and ground reinforcements, lighting, road signs, and wildlife barriers. Future studies should pay more attention to those components of the network that contain concrete.

Since the present study is limited to roads, the three main road materials are considered: asphalt, steel, and aggregates (including gravels used for gravel roads and sand and stone used in the base layer of roads). Steel is mainly used in guard rails. The reason for including the guard rails, even though their mass is small compared to asphalt and aggregates, is that steel has a significantly higher emission factor per kilogram compared to asphalt and gravel (Karlsson et al., 2020a). Concrete roads are excluded from the system boundary, as there are only 68 km of concrete roads in Sweden as of Year (2022) (VTI, 2022). Therefore, all paved roads included in the analysis are asphalt roads.

2.7. Material stock and flow analysis

A prospective bottom-up MFA of Swedish roads is performed to showcase the applicability of the proposed method. Furthermore, the embodied emissions associated with the prospective material flows are estimated using emission factors. To investigate the future embodied emissions of roads and the potential to reduce these emissions, two different future scenarios are constructed and analyzed. The method applied in each analysis step is described below, subsections 2.7.1-2.7.4.

2.7.1. Material stock estimation

A bottom-up estimation of the material stock is accomplished by summing the amounts of relevant materials that are present within the system boundary at a certain time (Gerst and Graedel, 2008). The flow of materials can be quantified based on the stock, using a stock-driven MFA approach (Müller, 2006). This stock in roads is quantified using the archetype-based approach introduced by Schiller (2007), and this specific approach has been adapted from Miatto et al. (2017). All roads are grouped based on archetype and width information into the following six categories: 1) one-lane road; 2) highway; 3) two-lane road wide; 4) two-lane road normal; 5) 2 + 1 road; and 6) gravel road. The material stock is calculated by multiplying the inventories of the roads by an MI factor, as expressed by **Equation (1)**:

$$MS_{road_{t,m}} = \sum_{i,j,r} L_{roadi,j,r,t} * W_{roadi,j,r,t} * MI_{roadi,m} \quad (1)$$

where $MS_{road_{t,m}}$ is the total mass of material stock of roads in year t of material m, $L_{roadi,j,r,t}$ is the length of road type i of road segment j in region r in year t in km, $W_{roadi,j,r,t}$ is the width (both actual and predicted) of road type i of road segment j in year t in km, and $MI_{roadi,m}$ is the material intensity ($tonne/m^2$) of material m for road type i.

2.7.2. Prospective material flow analysis

The prospective stock-driven MFA model for each year in the period of 2023–2045 is implemented in the Open Dynamic Material Systems

(ODYM) model (Pauliuk and Heeren, 2020). ODYM is an open-source software framework developed for dynamic MFA that involves multiple products and materials. The dynamic stock modeling sub-module was used to conduct the analysis.

The first step of the analysis is to calculate the materials need for the construction of new roads in each year ($M\_inflow\_new_{t+1}$), as expressed in **Equation (2)**:

$$M_{inflow_{new}\,t+1,m} = \sum_{i,r} L_{road\_proj_{i,r,t}} * W_{road\_proj_{i,r,t}} * MI_{road_{i,m}} \quad (2)$$

where $L_{road\_proj_{i,r,t+1}}$ is the projected length of new construction of road type i in region r at time $t+1$, and $W_{road\_proj_{i,r,t+1}}$ is the width of projected new construction of road type i in region r at time $t+\ 1$. The widths are assumed to be the average of all road widths of each road type i. For further information on the projection of new construction the reader is referred to the *Supplementary Information*.

The second step is to calculate the materials needed for the maintenance of roads. It is assumed that roads are not fully demolished, but instead are only maintained. At the end-of-life for each road section, only a fraction of the top asphalt layer is removed, and the road is subsequently repaved. This activity is defined as maintenance of the road. The assumption made is that each individual section of road is only maintained at the end-of-life with 50 mm of the asphalt layer being removed and subsequently repaved (based on information from experts in the road construction industry). Therefore, the frequency of maintenance depends on the lifetime of the road section.

Based on the stock and lifetime data, the amount of material needed for the maintenance of roads ($M_{inflow_{main}\,t,m,r}$) of material m in region r in year t is calculated using **Equation (3)**:

$$M_{inflow_{main}\,t,m,r} = \sum_{i,j} L_{road\,i,j,r,t} * W_{road\,i,j,r,t} * MI_{main\,i,m} - \sum_{\tau=t0}^{t-1} M_{inflow_{main}\,\tau,m,r} * Survival_{main\,t-\tau,r} \quad (3)$$

where $L_{road\,i,j,r,t}$ is the length of road type i of road segment j in region r in year t in km, $W_{road\,i,j,r,t}$ is the width (both actual and predicted) of road type i of road segment j in year t in km, and $Survival_{main\,t-\tau,r}$ is the survival curve using complementary cumulative distribution function of roads in year t in region r based on the Weibull distributions outlined in Table 3. The MIs for maintenance flow $MI\_main_{i,t,m}$ for road type i for material type m were used.

The total inflow of material ($M\_inflow\ \_total_{t,m}$) is, thus, the sum of the material inflows for new constructions and maintenance, as expressed by **Equation (4)**:

$$M_{inflow_{total}\,t,m} = M_{inflow_{new}\,t,m} + \sum_{r} M_{inflow_{main}\,t,m,r} \quad (4)$$

2.7.3. Embodied emissions

The embodied emissions in roads are calculated using **Equation (5)**:

$$Embodied_{emission\,t} = \sum_{i,t,m} M_{inflow_{total}\,i,t,m} * EF_m \quad (5)$$

where $Embodied\_emission_t$ is the total embodied CO_2 emissions for all Swedish roads in year t, $M\_inflow\ \_total_{t,m}$ is the total inflow of material from **Equation (5)**, and EF_m is the emission factor for material m. The emission factors (EF) used here are based on estimates made by Karlsson et al. (2020b).

2.7.4. Scenario analysis

A scenario analysis is carried out to assess the potential for reducing embodied emissions. The scenarios explore different developments in material production and new construction versus maintenance activities. Two different sets of emission factors (*EFs*) were used for the future projection scenarios (See Table 1 in the *Supplementary Information*). In the *Business-as-usual* scenario, no emissions reductions are assumed to be achieved in relation to materials production from Year (2020) onwards. In the *Emission reduction* scenario, significant emissions reductions are assumed to be achieved in relation to basic materials production over the studied period. The purpose of this scenario is to establish a baseline for comparison. For further information on the scenarios, the *EFs,* and the projection of new construction the reader is referred to the *Supplementary Information*.

3. Results

3.1. Machine learning model selection

The performance levels of the trained models are shown in Table 4. The performance levels of all four models are similar, with the biggest difference seen for the training computation time. Despite the relatively low computation time, the training process for machine learning models still requires significant computational power. The MAE can be interpreted as the average of the absolute prediction errors for each road section width (in meters).

The results from the LightGBM and XGBoost models are very similar, with XGBoost producing a slightly lower MAE, MAPE, and higher R^2 value. Both LightGBM and XGBoost are implemented to have automatic multithreading enabled, so as to achieve shorter computation time by parallelizing the computation on all available computer cores simultaneously. The similarity in the results is not unexpected, as the two models are variations of the gradient-boosting algorithm.

The RF algorithm has fewer hyperparameters available for tuning and, therefore, performs slightly poorer in terms of the MAE, MAPE and R^2 parameters. Unlike the LightGBM and XGBoost models, the RF model does not implement approximated training algorithms, and this results in a longer computation time. Approximated training algorithms are designed to speed up computation by building histograms for each feature value and performing iterations through the histograms rather than the real dataset. The CatBoost model results in the highest MAE and MAPE values and the lowest R^2 value of all four tested algorithms. This is most likely due to the relatively low number of categorical features (6) in the dataset, which does not fully take advantage of the algorithm's implementations on categorical features. All four chosen models are tree-based because such an algorithm reduces overfitting; other models

Table 4
Evaluation metrics for trained models, including computational time.

Algorithm	MAE	MAPE (%)	R^2	Computation time[a]
Random forest (RF)	0.623	13.4	0.748	3 min 22 s
XGBoost	0.567	12.5	0.784	30 s
CatBoost	0.669	14.4	0.728	2 min 10 s
LightGBM	0.576	12.6	0.781	32 s

MAE, mean absolute error; MAPE, mean absolute percentage error; R^2, coefficient of determination.

[a] Computed using a desktop PC with 12 CPU cores.

such as deep learning (Lecun et al., 2015) could be tested in future work.

Fig. 3 provides a visualization of the results of the XGBoost model. Actual values (blue) refer to the validation dataset used in the training process and the predicted values (green) are the results produced by the XGBoost model using the same training dataset. The results show that the model is less capable of predicting extreme values, especially road widths exceeding 10 m. Furthermore, the model under predicts roads with 6 m in width and results in a normalized distribution for road widths between 4.5 and 8 m. Overall, the model does capture the distribution of the real data relatively well in terms of filling in missing data.

Thus, the overall best-performing model selected is the XGBoost model. The subset of data with missing road width is predicted by the XGBoost models using the same set of training features. The predicted data are subsequently appended to the subset of data with existing width data to create a hybrid road attributes dataset. This dataset is subsequently used to estimate the material stock, material flows, and embodied emissions of the Swedish roads.

3.2. Material stock

Fig. 4 provides a comparison of the Swedish road sections based on the type of road ownership with respect to road lengths and road area (top plots) and the material stock divided according to road type and material type (bottom plots). In total, there are 638,632 km of paved and gravel roads in Sweden. Privately owned roads have the longest absolute length, being approximately four-times longer than the state-owned roads and 10-times longer than the roads owned by the municipalities. Gravel roads represent 92% of the total road length for privately owned roads. Taking the width and the road area into account, the differences between privately owned roads and other ownership type are diminished, with the total area of privately owned roads being 272% larger than that of the stated-owned roads and 600% larger than the area of the municipally owned roads. In general, the municipalities own roads in urban areas and, thus, have the lowest number of roads in terms of length and area.

In total, there are 2011 Mton (10^6 tonnes) of in-use material stock in all Swedish roads (bottom panels, ***Fig. 4***). The privately owned, municipally owned, and state-owned roads represent 56.7%, 12.2%, and 31.1% of the material stock, respectively. The share of in-use material stock in privately owned roads is much smaller than the share of road length because gravel roads have the lowest MI per km^2. Despite this lower share, privately and municipally owned roads constitute almost 70% of the total in-use stock, which highlights the need to have more accurate data for these roads. Fig. 4*c* shows the material stock for each road owner distinguished by road type. Two-lane roads represent the highest share of in-use material stock at 60.3%, and the majority of the two-lane roads are owned by the state. Fig. 4*d* shows the material composition of the in-use stock for each road owner. The aggregated values represent the highest share of in-use material stock for all road owners in terms of mass, at 96.9%, 92.2%, and 92.3% for private owners, municipalities, and the state, respectively. Most of the aggregates are used to form the base course and unbound layer and are not removed during the maintenance process. As steel is used only in guard rails in the system boundary, it accounts for very low shares of the stock at 0.0003%, 0.0002%, and 0.0458% for the private, municipal and state owners of roads, respectively. State-owned roads have the largest share of steel due to the higher level of steel usage in highways and 2 + 1 roads.

3.3. Material flows

Fig. 5 presents the annual inflows of aggregates and asphalt (Fig. 5a) and for steel (Fig. 5b) for the period of 2020–2045, as obtained from the stock-driven MFA. The trends regarding the inflows of all three materials are very similar because it is assumed that the road surface and the guard rails will be maintained at the same time. The average inflow for asphalt is 11.84 Mt and the average inflow for steel is 0.15 Mt, and for each year 0.28 Mt of asphalt and 1394 tonnes of steel are used for new construction. The peak in inflows in 2022 is due to the right-skewed Weibull distribution and since the same lifetime is used for all three materials the peaks happen at the same time.

3.4. Embodied emissions

Fig. 6 provides the results of the scenario-based embodied emissions analysis. In both scenarios, a constant level of new construction is assumed. The variations in the embodied emissions each year in the *Business-as-usual* scenario are caused by the lifetime in the MFA model, as different numbers of roads are maintained each year. Panels a and b in Fig. 6 show the yearly embodied emissions for both scenarios divided according to the three materials used. In the *Emission Reduction* scenario, the embodied emissions decrease from 0.7 Mt in 2020 to 0.34 Mt in 2045, which represents a 51.4% reduction. Despite only representing 0.024% of the total in-use stock by mass, steel contributes large shares of the embodied emissions in Year (2045): 40% in the *Business-as-usual*

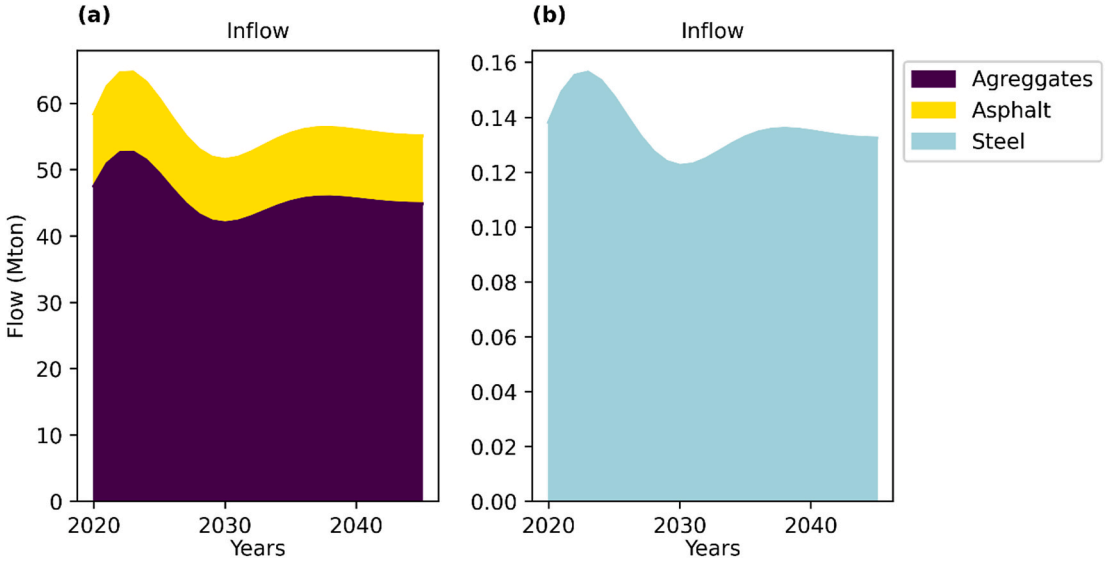

Fig. 5. Inflows in Mt (10^6 tonnes) of aggregates, asphalt, and steel assuming a constant rate of new construction. a) aggregates and asphalt. b) steel.

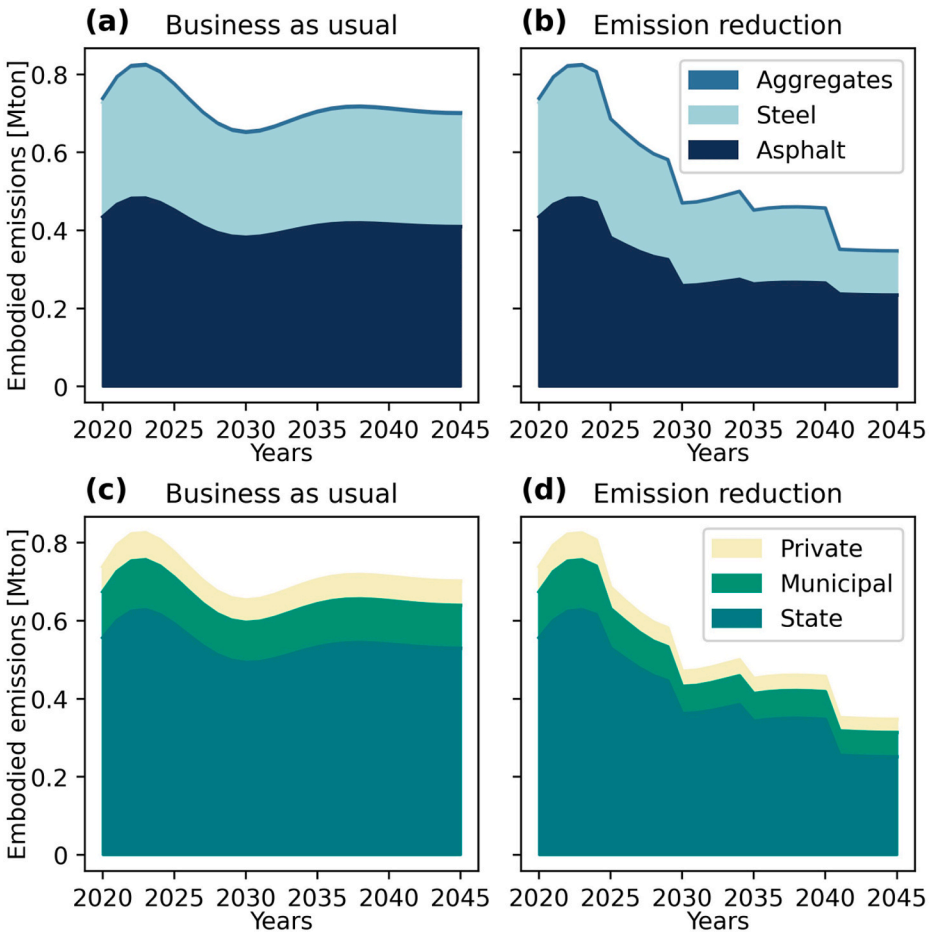

Fig. 6. Annual embodied emissions in Mt (10^6) for the period of 2020–2045: a) *business-as-usual scenario*, distinguished according to types of materials,; b) *emissions reduction scenario*, distinguished according to types of materials, in Mt; c) *business-as-usual scenario*, distinguished according to types of road ownership; d) *emissions reduction scenario*, distinguished according to types of road ownership.

scenario; and 32.4% in the *Emission reduction* scenario. This is because the emission factor for steel is many magnitudes greater than those of asphalt and aggregates. An assumption is made that low-carbon steel (for the type of steel used in construction) will not come onto the market before Year 2045. Asphalt contributes the largest share of embodied emissions, at 58.6% for the *Business-as-usual* scenario and 67.4% for the *Emission reduction* scenario in the year 2045.

Panels c and d in Fig. 6 show the embodied emissions for the two scenarios, distinguishing between the different road owners. These subplots demonstrate that privately owned roads contribute a significant share of the embodied emissions, and that greater attention needs to be paid to data gathering processes and to policy-making that targets this segment of the road system. State-owned roads contribute the largest share of embodied emissions at 75.7% for the *Business-as-usual* scenario and 73.5% for the emissions reduction scenario in Year (2045). Municipally owned and privately owned roads contribute 15.7% and 8.6%, respectively, in the *Business-as-usual* scenario, and 17.6% and 8.8%, respectively, in the emissions reduction scenario for Year 2045. This highlights that to reach overall net-zero embodied emissions, non-state road owners need to be included in the policy-making process.

We compare the results of the material flow analysis and embodied emissions to other studies to contextualize the results. Wiedenhofer et al. (UN Environment Programme, 2019) assessed the material stock, flow, and embodied emissions of all mobility infrastructure (roads, railways, bridges, tunnels, etc.) for 180 countries using a similar spatially explicit, bottom-up approach. The stock and flow and emissions result provided by Wiedenhofer et al. were grouped into minimum, mean, and maximum values. The road length in our dataset is 638,632 km while the total length in Wiedenhofer et al. is 542,575 km. The maximum total material stock for all roads in Sweden from Wiedenhofer et al. is 2338.5 Mton while our results is 2011 Mton. This is likely due to the difference in road width as our study does not assume a constant width for each road type. Our asphalt inflow is closer to Wiedenhofer et al.'s minimum inflow (11.65 Mt vs 10.88 Mt) while our aggregates inflow is closer to the maximum inflow (50.89 Mt vs 58.94 Mt). The differences in the estimates of aggregate inflows are expected since our analysis mainly focused on predicting width of private and municipal roads which is largely consisted of gravel roads.

Furthermore, we compare our results using widths for each road section to an MFA model that has the same input besides using assumed average road widths for each type of roads. The assumed average road width is taken from NVDB. The results of this analysis show that the material stock from the average road width model is 8.62%, 12.0%, and 37.2% larger respectively for aggregate, asphalt, and steel respectively. This corresponds to a 24.3% and 22.2% increase in embodied emissions in the year 2045 in the Business as usual and Emissions reduction scenario respectively. The difference in results could be due to the uncertainty introduced by the machine learning model, but it does highlight the need for more granular data for material stock and embodied analysis of roads.

4. Discussion

4.1. Applicability of machine learning methods for predicting road width

Machine learning models, and more specifically gradient-boosting

algorithms, perform well in predicting road width in situations where data are missing from the statistical dataset, with the best-performing model achieving a R^2 value of 0.784. A key challenge in predicting the geometric attributes of municipal and private roads is that detailed data, such as layer thickness and traffic flow, are not available and are also unlikely to be available for other countries. In machine learning workflows, the training dataset and the prediction dataset must have the same set of features. This means that features that are likely to have strong correlations with the geometric attributes cannot be included in the model. The results from this work demonstrate that non-physical features, such as network features, do correlate with physical road attributes, and that machine learning models are able to capture complex relationships so as to facilitate predictions. Future work could investigate the possibility to utilize a more-urban morphological network-based analysis for MFA studies. Furthermore, deep learning approaches could be tested and compared with gradient-boosting machine learning algorithms (Lecun et al., 2015).

Another limitation associated with the application of the method introduced in this work is the availability of existing data. Although many types of data are publicly available in Sweden, there remain large gaps in the dataset. In countries with lower availability of data, the potential for machine learning to compensate for missing data will be more limited. In cases where machine learning cannot be applied due to lack of available data, the data must first be gathered, for example using remote sensing from satellites or night-time light approaches. Such methods might also be more appropriate for developing countries with limited available data.

4.2. Generalizability of the approach

The proposed framework of combining machine learning and MFA to quantify embodied carbon emissions using an incomplete data set is highly generalizable given available data. All the generated features utilize basic GIS data such as road and building footprints. The use of GIS data is essential for this approach as most of the features are generated using spatial analysis tools. The strength of machine learning models is that they perform ("learn") differently depending on the specific dataset. Thus, it is possible to use a different set of features depending on the context and data availability. This study demonstrates the possibility of using a very limited number of physical attributes of the roads as features (length and surface type) and still achieves a good accuracy (R^2 value). One potential future application of this approach is to OpenStreetMap where the road network attributes are incomplete (Herfort et al., 2022). The strength of using machine learning to fill-in the gap in the data is that the data requirement is flexible and can be varied based on the research question. Furthermore, the proposed approach can also be used to predict other physical road attributes.

4.3. Embodied emissions from new construction versus maintenance

To achieve the goal of net-zero embodied emissions, a potential policy could be to scale down or even completely halt the construction of new roads. For example, in Wales, all new major road construction projects have been scrapped and any new construction projects must not increase the levels of carbon emissions (BBC), (The Guardian). This reduction in new construction could in the future be the result of decreased driving demand through a combination of technological advancements, such as self-driving cars and behavioral changes (Morfeldt et al., 2023). This could reduce the levels of both embodied emissions and material consumption. Fig. 7 shows the embodied emissions, divided between maintenance and new construction for the two scenarios. In both scenarios, new constructions contribute only around 2.9% of the total annual embodied emissions. The system boundary does not include other transport infrastructures that use concrete, such as bridges and tunnels; the percentage of embodied emissions resulting from new construction will be higher if other infrastructures are included (Karlsson et al., 2020a).

As Sweden is a developed country with well-established infrastructures, the annual addition of new roads is small relative to the existing stock. Therefore, the largest shares of material consumption and embodied emissions are from road maintenance activities. This indicates that there are two main pathways to reducing embodied emissions: 1) a better maintenance regime to prolong the road lifetime; and 2) procurement policies that facilitate the production of low-embodied-emissions materials. In contrast, the demand for new road construction remains high in developing economies, where the emissions associated with providing road access to at least 97.5% of the population in any country are estimated to be about 2 $GtCO_2$ (Wenz et al., 2020). Therefore, measures that would reduce the levels of embodied emissions for new construction of roads, such as policies for material efficiency, should be promoted in developing countries (Hertwich et al., 2020).

Since the maintenance of roads contributes the largest share of embodied emissions, improved maintenance regimes, such as predictive maintenance, could prolong the lifetimes of roads and, thereby, reduce the amount of material required for maintenance of the Swedish road network. This should be applicable to other developed countries with well-developed road networks, such the US where an estimated 75% of the yearly material inflow into the road network material stock is linked

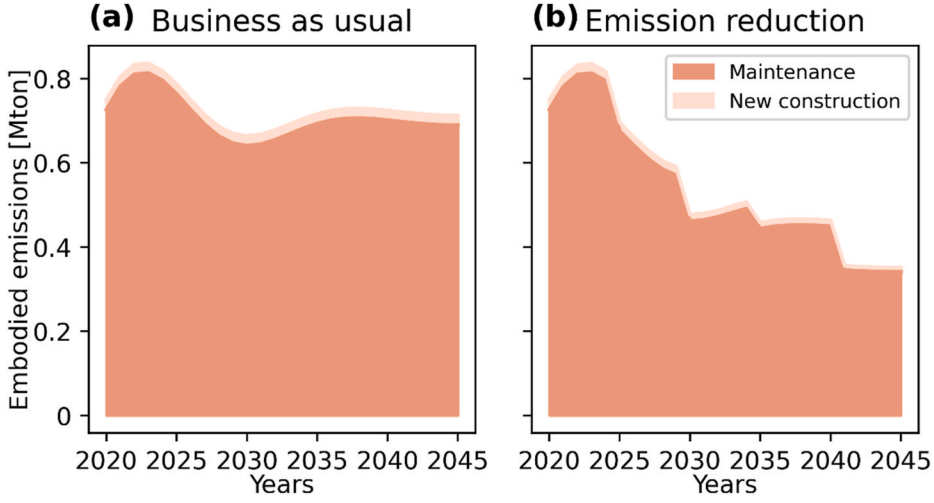

Fig. 7. Embodied emissions in Mt (10^6 tonnes) distinguishing between maintenance and new construction activities from Year (2020) to Year 2045. a) *Business-as-usual* scenario. b) *Emission reduction* scenario.

to maintenance activities (Miatto et al., 2017). On the other hand, the implementation of embodied carbon emission requirements for materials used during maintenance and new construction could be an effective policy for emissions reductions. A well-implemented procurement requirement and emissions declaration can lower the technological and market barriers by creating a market for low-CO_2 products, as discussed by Löfgren & Rootzén (Löfgren et al., 2021). This could be especially effective for asphalt, since road construction is the main consumer of asphalt.

4.4. Limitations and future work

The focus of the present work is the development of methods to fill the gaps in the road dataset. Thus, the MFA used the more-aggregated archetype approach. As previously discussed, a layer thickness-based approach can produce more-accurate stock and flow results. The method introduced in this work could be used to predict other geometric attributes such as layer thickness, although the challenge is that each prediction introduces uncertainty and using a predicted value for further predictions simply increases the level of uncertainty related to the results. An alternative could be to apply an archetype-based method to estimate pavement thickness, such as that developed by Wang et al. (2022), and to combine this with the method proposed in the present work, thereby creating a hybrid model.

Furthermore, the lifetimes used in the MFA are static and do not consider potential future changes in climatic conditions or traffic flows that might affect the lifetimes of the roads. The inclusion of future climate change scenarios may introduce significant uncertainties related to the maintenance needs of roads (Valle et al., 2017), (Guest et al., 2020). Thus, the results of the analysis represent a snapshot of potential future material flows and embodied emissions, assuming all factors remain constant.

For the machine learning models, future work should aim to improve the interpretability of the machine learning results by including feature importance analysis. This is outside the scope of this study as the focus is to demonstrate the applicability of machine learning to complement statistical data. In the context of Industrial Ecology, the ease with which machine learning models can be interpreted is especially important in studies that investigate the relationships between cause and outcomes, which could be used to inform decision-making processes (Donati et al., 2022). For a comprehensive review of the different explainability assessment methods of machine learning models, the reader is referred to the paper of Ali et al. (2023).

Material recycling was excluded from the analysis due to the lack of information on road surface type and layer thickness data for municipally owned and privately owned roads. If modeling recycling in greater detail, the reduction in embodied emissions and increase in material efficiency might be higher than those shown in this work. Along with the development of various data collection methods and machine learning models, future work should consider the trade-off between the complexity of the chosen modeling method and the magnitudes of uncertainties associated with the method.

Lastly, the spatial aspect of the proposed approach should be further explored. For example, this approach can be expanded to analyze the concept of circularity hub of construction materials as describe by Tsui et al. (2024). The spatial aspect of roads can be used to calculate the transport distances of a potential circularity hub. Furthermore, the material outflows from the roads can also be considered for potential circularity hubs. Similarly, the material stock efficiency of stock as shown in Wang et al. (2022) can be expanded to include municipal and privately owned roads to gain a better understanding of stock efficiency for rural areas.

5. Conclusion

In the coming decades, decarbonization of transport infrastructures will be a major challenge in relation to limiting global warming. Understanding the future material flows involved in the construction and maintenance of the transport infrastructure is essential for meeting ambitious emission reduction goals. While detailed datasets with high geographic resolution and physical attributes are crucial, datasets are usually unavailable. This study proposes a machine learning-based method for predicting missing statistical data regarding physical attributes in a road dataset, which can be used for estimating material flows and embodied emissions. Sweden is used as a case study, since the road dataset is publicly available. Four machine learning algorithms are tested, and the best-performing model is chosen to predict the missing road width data. The predicted road widths are used to complete the dataset, and a stock-driven MSFA is conducted for the period of2020–2045. The material flows are thereafter used to calculate scenario-based embodied emissions.

Machine learning models give good prediction results, and the best-performing model is called XGBoost. The material stock estimation shows that non-state-owned roads contribute 47.5% of the material stock, which is the main source of the missing width data. The emissions reduction scenario shows that by Year (2045), the yearly embodied emissions for Swedish roads can be reduced by 51%. Furthermore, new construction contributes only 2% of the total yearly embodied emissions. This points to the importance of improved road maintenance regimes and embodied carbon emissions requirements for the materials used in road maintenance, as well as new strategies to reduce the embodied emissions of roads. While the scenario results are relevant to the Swedish context, the methodological approach is equally applicable to other countries, provided that the requisite underlying data are available.

In terms of future research efforts, the application of machine learning models can be expanded to predict other physical attributes of roads, such as layer thickness. Furthermore, the feature importance of the machine learning models can be included to improve the explainability of results. If more data become available, more-detailed models of layer wear and tear and material recycling should be included to investigate the embodied emissions reduction potential and material efficiency.

CRediT authorship contribution statement

Qiyu Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Johan Rootzén:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition. **Filip Johnsson:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was financed by Mistra Carbon Exit research programme. The authors acknowledge and are thankful for the discussion and comments from Susanna Toller, Sofiia Miliutenko, Carolina Liljenström and Stefan Uppenberg.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cesys.2024.100211.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: a next-generation hyperparameter optimization framework. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2623–2631. https://doi.org/10.1145/3292500.3330701.

Ali, S., et al., 2023. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf. Fusion 99 (March), 101805. https://doi.org/10.1016/j.inffus.2023.101805.

Augiseau, V., Barles, S., 2017a. Studying construction materials flows and stock: a review. Resour. Conserv. Recycl. 123, 153–164. https://doi.org/10.1016/j.resconrec.2016.09.002.

Augiseau, V., Barles, S., 2017b. Studying construction materials flows and stock: a review. Resour. Conserv. Recycl. 123, 153–164. https://doi.org/10.1016/j.resconrec.2016.09.002.

Bao, Y., Huang, Z., Wang, H., Yin, G., Zhou, X., Gao, Y., 2023. High-resolution quantification of building stock using multi-source remote sensing imagery and deep learning. J. Ind. Ecol. 27 (1), 350–361. https://doi.org/10.1111/jiec.13356.

BBC, "All major road building projects in Wales are scrapped." https://www.bbc.com/news/uk-wales-64640215.

Brons, M., Dijkstra, L., Ibáñez, J.N., Poelman, H., 2022. Road Infrastructure in Europe Road Length and its Impact on Road Performance. https://doi.org/10.2776/21558.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Currie, P.K., Musango, J.K., May, N.D., 2017. Urban metabolism: a review with reference to Cape Town. Cities 70, 91–110. https://doi.org/10.1016/j.cities.2017.06.005.

Deng, T., Fu, C., Zhang, Y., 2022. What is the connection of urban material stock and socioeconomic factors? A case study in Chinese cities. Resour. Conserv. Recycl. 185, 106494 https://doi.org/10.1016/j.resconrec.2022.106494.

Donati, F., et al., 2022. The future of artificial intelligence in the context of industrial ecology. J. Ind. Ecol. https://doi.org/10.1111/jiec.13313 jiec.13313.

Ebrahimi, B., Rosado, L., Wallbaum, H., 2022. Machine learning-based stocks and flows modeling of road infrastructure. J. Ind. Ecol. 26 (1), 44–57.

Eurostat, 2022. Eurostat. https://ec.europa.eu/eurostat/databrowser/view/RAIL_IF_LINE_TR_custom_3356742/default/table?lang=en.

Fleischmann, M., 2019. Momepy: urban morphology measuring toolkit. J. Open Source Softw. 4 (43), 1807. https://doi.org/10.21105/joss.01807.

Funke, S., Schirrmeister, R., Storandt, S., 2015. Automatic extrapolation of missing road network data in OpenStreetMap. CEUR Workshop Proc. 1392 (January), 27–35.

Gerst, M.D., Graedel, T.E., 2008. In-use stocks of metals: status and implications. Environ. Sci. Technol. 42 (19), 7038–7045. https://doi.org/10.1021/es800420p.

Gontia, P., Thuvander, L., Ebrahimi, B., Vinas, V., Rosado, L., Wallbaum, H., 2019. Spatial analysis of urban material stock with clustering algorithms: a northern European case study. J. Ind. Ecol. 23 (6), 1328–1343. https://doi.org/10.1111/jiec.12939.

Guest, G., Zhang, J., Maadani, O., Shirkhani, H., 2020. Incorporating the impacts of climate change into infrastructure life cycle assessments: a case study of pavement service life performance. J. Ind. Ecol. 24 (2), 356–368. https://doi.org/10.1111/jiec.12915.

Guo, Z., Hu, D., Zhang, F., Huang, G., Xiao, Q., 2014. An integrated material metabolism model for stocks of urban road system in Beijing, China. Sci. Total Environ. 470–471, 883–894. https://doi.org/10.1016/j.scitotenv.2013.10.041.

Guo, Z., Shi, H., Zhang, P., Chi, Y., Feng, A., 2017. Material metabolism and lifecycle impact assessment towards sustainable resource management: a case study of the highway infrastructural system in shandong peninsula, China. J. Clean. Prod. 153, 195–208. https://doi.org/10.1016/j.jclepro.2017.03.194.

Han, J., Xiang, W.-N., 2013. Analysis of material stock accumulation in China's infrastructure and its regional disparity. Sustain. Sci. 12.

Herfort, B., Lautenbach, S., Porto De Albuquerque, J., Anderson, J., Zipf, A., Ao Porto De Albuquerque, J., 2022. Investigating the Digital Divide in OpenStreetMap: Spatio-Temporal Analysis of Inequalities in Global Urban Building Completeness, pp. 1–14. https://doi.org/10.1038/s41467-023-39698-6.

Hertwich, E., et al., 2020. Resource Efficiency and Climate Change: Material Efficiency Strategies for a Low-Carbon Future. https://doi.org/10.5281/zenodo.3542680.

IPCC, 2022. *Climate Change 2022: Impacts, Adaptation, and Vulnerability*. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, p. 3056. https://doi.org/10.1017/9781009325844.

Jiang, B., Claramunt, C., 2004. Topological analysis of urban street networks. Environ. Plann. Plann. Des. 31 (1), 151–162. https://doi.org/10.1068/b306.

Jordahl, K., et al., 2019. Geopandas/Geopandas: V0. 6.0. Zenodo.

Karlsson, I., Rootzén, J., Johnsson, F., 2020a. Reaching net-zero carbon emissions in construction supply chains – analysis of a Swedish road construction project. Renew. Sustain. Energy Rev. 120, 109651 https://doi.org/10.1016/j.rser.2019.109651.

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

Kasraian, D., Maat, K., Stead, D., van Wee, B., 2016. Long-term impacts of transport infrastructure networks on land-use change: an international review of empirical studies. Transport Rev. 36 (6), 772–792. https://doi.org/10.1080/01441647.2016.1168887.

Ke, G., et al., 2017. LightGBM: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 2017-Decem, 3147–3155. Nips.

Khumvongsa, K., Guo, J., Theepharaksapan, S., Shirakawa, H., Tanikawa, H., 2023. Uncovering urban transportation infrastructure expansion and sustainability challenge in Bangkok: insights from a material stock perspective. J. Ind. Ecol. https://doi.org/10.1111/jiec.13342 jiec.13342.

Kloostra, B., Makarchuk, B., Saxe, S., 2022. Bottom-up estimation of material stocks and flows in Toronto's road network. J. Ind. Ecol. 26 (3), 875–890.

Kropf, K., 2018. The Handbook of Urban Morphology. John Wiley & Sons.

Lanau, M., Liu, G., 2020. Developing an urban resource cadaster for circular economy: a case of odense, Denmark. Environ. Sci. Technol. 54 (7), 4675–4685. https://doi.org/10.1021/acs.est.9b07749.

Lanau, M., et al., 2019. Taking stock of built environment stock studies: progress and prospects. Environ. Sci. Technol. 53 (15), 8499–8515. https://doi.org/10.1021/acs.est.8b06652.

Lantmäteriet, 2023a. Building download, INSPIRE. Lantmateriet.se. https://www.lantmateriet.se/en/geodata/geodata-products/product-list/buildning-download-inspire/.

Lantmäteriet, 2023b. Real property register. Lantmateriet.se. https://www.lantmateriet.se/en/real-property/property-information-/real-property-register/.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Liljenström, C., Toller, S., Åkerman, J., Björklund, A., 2019. Annual climate impact and primary energy use of Swedish transport infrastructure. Erratum Published in EJTIR 20 (2), 36–40, 40.

Löfgren, Å., Rootzén, J., 2021. Brick by brick: governing industry decarbonization in the face of uncertainty and risk. Environ. Innov. Soc. Transit. 40, 189–202. https://doi.org/10.1016/j.eist.2021.07.002.

Meijer, J.R., Huijbregts, M.A.J., Schotten, K.C.G.J., Schipper, A.M., 2018. Global patterns of current and future road infrastructure. Environ. Res. Lett. 13 (6) https://doi.org/10.1088/1748-9326/aabd42.

Miatto, A., Schandl, H., Wiedenhofer, D., Krausmann, F., Tanikawa, H., 2017. Modeling material flows and stocks of the road network in the United States 1905–2015. Resour. Conserv. Recycl. 127, 168–178. https://doi.org/10.1016/j.resconrec.2017.08.024.

Miatto, A., Dawson, D., Nguyen, P.D., Kanaoka, K.S., Tanikawa, H., 2021. The urbanisation-environment conflict: insights from material stock and productivity of transport infrastructure in hanoi, vietnam. J. Environ. Manag. 294, 113007 https://doi.org/10.1016/j.jenvman.2021.113007.

Morfeldt, J., et al., 2023. Emission pathways and mitigation options for achieving consumption-based climate targets in Sweden. Commun. Earth Environ. 4 (1), 1–14. https://doi.org/10.1038/s43247-023-01012-z, 2023 41.

Müller, D.B., 2006. Stock dynamics for forecasting material flows—case study for housing in The Netherlands. Ecol. Econ. 59 (1), 142–156. https://doi.org/10.1016/j.ecolecon.2005.09.025.

Müller, D.B., et al., 2013. Carbon emissions of infrastructure development. Environ. Sci. Technol. 47 (20), 11739–11746. https://doi.org/10.1021/es402618m.

Nasir, U., Chang, R., Omrany, H., 2021. Calculation methods for construction material stocks: a systematic review. Appl. Sci. 11 (14), 6612. https://doi.org/10.3390/app11146612.

Nguyen, T.C., Fishman, T., Miatto, A., Tanikawa, H., 2019. Estimating the material stock of roads: the Vietnamese case study. J. Ind. Ecol. 23 (3), 663–673. https://doi.org/10.1111/jiec.12773.

Nilsson, J.-E., Svensson, K., Haraldsson, M., 2020. Estimating the marginal costs of road wear. Transport. Res. Part A Policy Pract. 139, 455–471. https://doi.org/10.1016/j.tra.2020.07.013.

Pauliuk, S., Heeren, N., 2020. ODYM—an open software framework for studying dynamic material systems: principles, implementation, and data structures. J. Ind. Ecol. 24 (3), 446–458. https://doi.org/10.1111/jiec.12952.

Pedregosa, F., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Persson, J., 2020. Sweden's Long-Term Strategy for Reducing Greenhouse Gas Emissions, p. 87.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. Catboost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 2018 (Section 4), 6638–6648. Decem.

Rockström, J., Gaffney, O., Rogelj, J., Meinshausen, M., Nakicenovic, N., Schellnhuber, H.J., 2017. A roadmap for rapid decarbonization. Science 355 (6331), 1269–1271. https://doi.org/10.1126/science.aah3443, 80-.

Rodrigue, J.-P., 2020. The Geography of Transport Systems. Routledge.

Rousseau, L.S.A., Kloostra, B., AzariJafari, H., Saxe, S., Gregory, J., Hertwich, E.G., 2022. Material stock and embodied greenhouse gas emissions of global and urban road pavement. Environ. Sci. Technol. 56 (24), 18050–18059. https://doi.org/10.1021/acs.est.2c05255.

Schiller, G., 2007. Urban infrastructure: challenges for resource efficiency in the building stock. Build. Res. Inf. 35 (4), 399–411. https://doi.org/10.1080/09613210701217171.

Statistikmyndigheten, 2023a. Statistics Sweden. Statistiska Centralbyrån. https://www.scb.se/en/.

Statistikmyndigheten, 2023b. DeSO – demografiska statistikområden. Statistiska Centralbyrån. https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/regionala-indelningar/deso—demografiska-statistikomraden/.

Statistikmyndigheten, 2023c. RegSO - regionala statistikområden. Statistiska Centralbyrån. https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/regionala-indelningar/regso—regionala-statistikomraden/.
Statistikmyndigheten, 2023d. Open data for grid statistics. https://www.scb.se/en/services/open-data-api/open-geodata/grid-statistics/.
Svenson, K., Li, Y., Macuchova, Z., Rönnegård, L., 2016. Evaluating needs of road maintenance in Sweden with the mixed proportional hazards model. Transp. Res. Rec. J. Transp. Res. Board 2589 (1), 51–58. https://doi.org/10.3141/2589-06.
Tanikawa, H., Hashimoto, S., 2009. Urban stock over time: spatial material stock analysis using 4d-GIS. Build. Res. Inf. 37 (5–6), 483–502. https://doi.org/10.1080/09613210903169394.
Tanikawa, H., Fishman, T., Okuoka, K., Sugimoto, K., 2015. The weight of society over time and space: a comprehensive account of the construction material stock of Japan, 1945-2010. J. Ind. Ecol. 19 (5), 778–791. https://doi.org/10.1111/jiec.12284.
The Guardian, "Welsh Road Building Projects Stopped after Failing Climate Review.".
Trafikverket, "Requirements for Reducing Greenhouse Gas Emissons," p. 2.
Trafikverket, 2022a. Lastkajen – sveriges väg- och järnvägsdata. Trafikverket. https://www.trafikverket.se/tjanster/data-kartor-och-geodatatjanster/hamta-var-oppna-data/lastkajen—sveriges-vag–och-jarnvagsdata/.
Trafikverket, 2022b. Klimatkalkyl – Infrastrukturens Klimatpåverkan Och Energianvändning I Ett Livscykelperspektiv. Trafikverket. https://www.trafikverket.se/for-dig-i-branschen/miljo—for-dig-i-branschen/energi-och-klimat/Klimatkalkyl/.
Tsui, T., et al., 2024. Spatial parameters for circular construction hubs: location criteria for a circular built environment. Circular Economy and Sustainability 4, 317–338. https://doi.org/10.1007/s43615-023-00285-y.
UN Environment Programme, 2019. Global Status Report for Building and Construction - towards a Zero-Emissions, Efficient and Resilient Buildings and Construction Sector.
Valle, O., Qiao, Y., Dave, E., Mo, W., 2017. Life cycle assessment of pavements under a changing climate. In: Pavement Life-Cycle Assess. - Proc. Pavement Life-Cycle Assess. *Symp. 2017*, pp. 241–250. https://doi.org/10.1201/9781315159324-25.
VTI, 2022. Betongvägar. https://www.vti.se/forskning/vag–och-banteknik/betongvagar.
Wang, Z., Wiedenhofer, D., Stephan, A., Perrotti, D., Van den bergh, W., Cao, Z., 2022. High-resolution mapping of material stocks in Belgian road infrastructure: material efficiency patterns, material recycling potentials, and greenhouse gas emissions reduction opportunities. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.2c08703.
Wenz, L., Weddige, U., Jakob, M., Steckel, J.C., 2020. Road to glory or highway to hell? Global road access and climate change mitigation. Environ. Res. Lett. 15 (7) https://doi.org/10.1088/1748-9326/ab858d.
Wiedenhofer, D., Steinberger, J.K., Eisenmenger, N., Haas, W., 2015. Maintenance and expansion: modeling material stocks and flows for residential buildings and transportation networks in the EU25. J. Ind. Ecol. 19 (4), 538–551. https://doi.org/10.1111/jiec.12216.
Yu, B., Li, L., Tian, X., Yu, Q., Liu, J., Wang, Q., 2021. Material stock quantification and environmental impact analysis of urban road systems. Transport. Res. Transport Environ. 93, 102756 https://doi.org/10.1016/j.trd.2021.102756.
Zhang, R., Yamashita, N., Liu, Z., Guo, J., Hiruta, Y., Shirakawa, H., 2023. Science of the Total Environment Paving the way to the future : mapping historical patterns and future trends of road material stock in Japan. Sci. Total Environ. 903 (August), 166632 https://doi.org/10.1016/j.scitotenv.2023.166632.
Zhao, F., Sun, H., Wu, J., Gao, Z., Liu, R., 2016. Analysis of road network pattern considering population distribution and central business district. PLoS One 11 (3), 1–17. https://doi.org/10.1371/journal.pone.0151676.
Zheng, A., Casari, A., 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc.