Predicting hazardous materials in the Swedish building stock using data mining

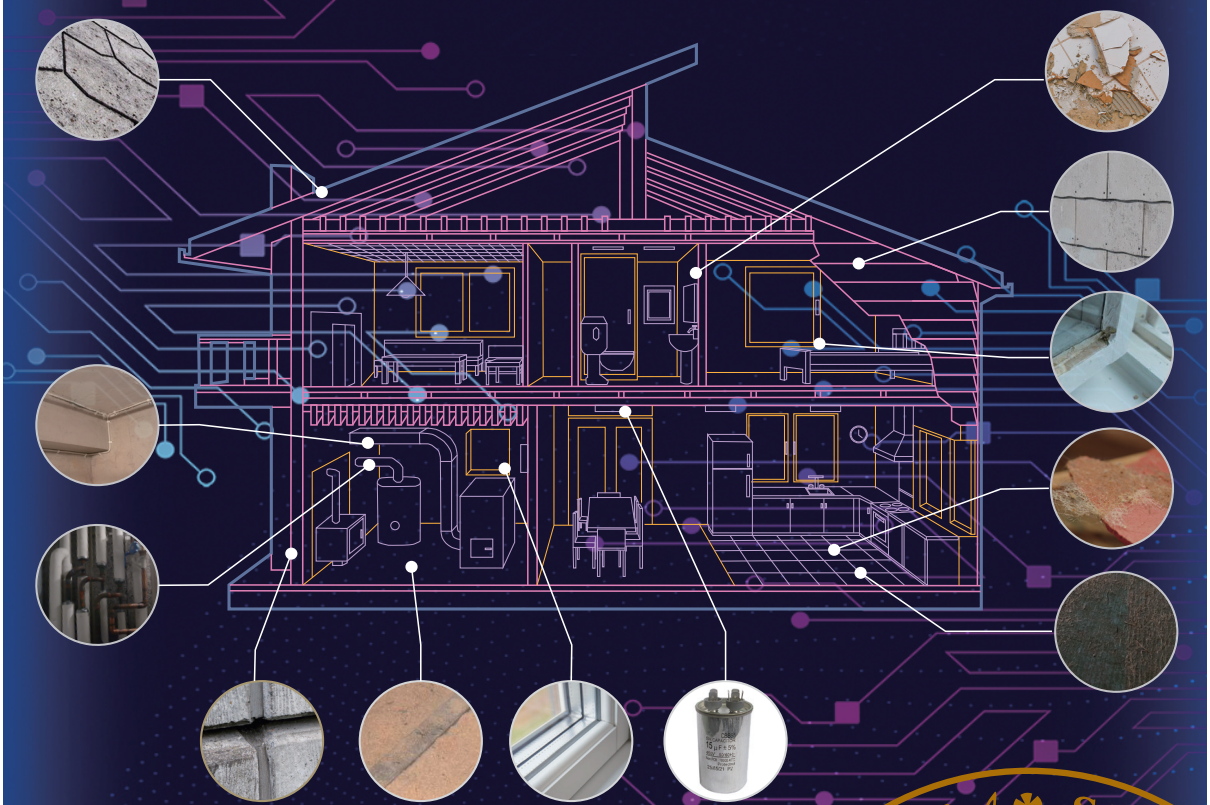
Wu, Pei-Yu

2022

Document Version:
Publisher's PDF, also known as Version of record

Predicting hazardous materials in the Swedish building stock using data mining

PEI-YU WU | FACULTY OF ENGINEERING | LUND UNIVERSITY



Printed by Media-Tryck, Lund 2022 NORDIC SWAN ECOLABEL 3041 0903

Transitioning into a circular construction is an inevitable trend to optimize resource efficiency. However, the presence of hazardous materials from the end-of-lifecycle buildings are incompatible with the ambition for a circular construction and challenges its realization in practice. Pre-demolition audits therefore act as a crucial means to assure quality of the recovered materials. Over years, these inventories of hazardous waste have been archived on a national scale, but are left out from building stock registers. What's their potential as input data for machine learning prediction? How can we leaverage the past detection records to trace the patterns of hazardous materials in the existing building stock. The thesis tries to answer these questions by mining the archived inventory data and information from relevant building registers. In search for emergent data-driven approaches for in situ hazardous material identification, the research front of construction and demolition waste management was presented. A promising hazardous material dataset and a machine learning pipeline were created as the means for assessing the potential detection and exposure risk. Also, the complexity of applied AI in addressing the diversity of building data is highlighted. The applied research aims to open a discussion for the necessity of establishing a standardized data collection infrastructure and assessment procedure to facilitate a data-driven hazardous material management.

The licentiate thesis by Pei-Yu Wu is a joint research work between Lund University and RISE Research Institutes of Sweden, funded by the Swedish Foundation for Strategic Research 2020-2023. Pei-Yu is a licensed architect and holds a master degree in design and construction project management. She is passionate for using building stock analysis to promote material recovery in circular economy. She believes that developing data-driven approaches can bring the vision to reality.

LUND UNIVERSITY

Faculty of Engineering
Department of Building and Environmental Technology
Division of Building Physics
ISBN 978-91-88722-77-5
ISRN LUTVDG/TVBH—22/1028—SE(110)
ISSN 0349-4950

Predicting hazardous materials in the Swedish building stock using data mining

Pei-Yu Wu

LUND
UNIVERSITY

LICENTIATE DISSERTATION
by due permission of the Faculty of Engineering, Lund University, Sweden.
To be defended at V:A, V-huset, John Ericssons väg 1, Lund.
February 24th, 2022 at 1.15 pm.

Faculty opponent
Ivo Martinac
Professor of Building Services and Energy Systems
KTH Royal Institute of Technology
Stockholm, Sweden

Organization	Document name
LUND UNIVERSITY Faculty of Engineering Department of Building and Environmental Technology Division of Building Physics	LICENTIATE DISSERTATION
	Date of issue February 24th, 2022
Author(s) Pei-Yu Wu	Sponsoring organization RISE Research Institutes of Sweden, The Swedish Foundation for Strategic Research

Title and subtitle

Predicting hazardous materials in the Swedish building stock using data mining

Abstract

Identifying the potential presence of hazardous materials can prevent unexpected decontamination costs and delays, as well as contaminant exposure in renovation and demolition work. However, the use of hazardous materials in past construction is comprehensive and lacks quantification. The current pre-demolition audit on the building basis is not efficient enough for large-scale mapping. As such, novel approaches for pattern identification need to be developed to facilitate contamination risk assessment in existing buildings. Data mining and its subfield machine learning present a new opportunity for using detection records to screen the likely presence of in situ hazardous materials in the national building stock.

The aim of the study is, therefore, to explore the potential of applied machine learning for predicting hazardous materials using building registers as input data and hazardous waste inventories as training and validation data. Considerable efforts have been dedicated to reducing the data uncertainty in merging and matching empirical data and building registers. The workflows of constructing a hazardous material dataset and a machine learning pipeline highlighted the complexity of processing unstructured, heterogeneous building-specific data. The results indicated that machine learning techniques succeed in characterizing suspected hazardous building materials, which is of significance for realizing the EU Construction and Demolition Waste Management Protocol. The detection likelihood of asbestos, PCB, CFC, and mercury were estimated according to inventory document types and building classes. Considering the building stock's diversity, a cross-validation matrix evaluating the quality and quantity of data subgroups was created for data stratification. Asbestos and PCB-containing materials in multifamily houses, schools, and commercial buildings were potential for modeling. Six supervised algorithms were used to test the prediction possibility. The average validation accuracies are 74% and 83% for predicting asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings.

Influential features to the prediction results were also visualized for expert knowledge interpretation, which has a practical implementation for assisting decision-making in constructing clean material loops. Construction year, floor area, and the number of stairwells and floors were influential for asbestos pipe insulation prediction, while construction year, balanced ventilation system, floor area are critical for PCB joints or sealants prediction. The proposed bottom-up modeling approach can potentially be scaled up to a national level and replicated in other countries with similar access to building stock data.

Keywords

Hazardous materials, Pre-demolition audits, Construction and Demolition waste, Swedish building stock, Data mining, Methodology, Machine learning, Prediction, Risk assessment

Classification system and/or index terms (if any)

Supplementary bibliographical information ISRN LUTVDG/TVBH—22/1028—SE(110)		**Language** English
ISSN and key title 0349-4950		**ISBN** 978-91-88722-77-5
Recipient's notes	**Number of pages** 110	Price
	Security classification	

Signature Date 2022-01-14

Predicting hazardous materials in the Swedish building stock using data mining

Pei-Yu Wu

LUND
UNIVERSITY

Cover photo by Pei-Yu Wu

Faculty of Engineering
Department of Building and Environmental Technology

ISBN 978-91-88722-78-2 (PDF)
ISBN 978-91-88722-77-5 (print)
ISSN 0349-4950

Printed in Sweden by Media-Tryck, Lund University
Lund 2022

The ambitions for a circular economy are high and unambiguous, but day-to-day experience shows that the transition still has many difficulties to overcome.

Bodar, Spijker and Lijzen et al.

Table of Contents

Abstract

Identifying the potential presence of hazardous materials can prevent unexpected decontamination costs and delays, as well as contaminant exposure in renovation and demolition work. However, the use of hazardous materials in past construction is comprehensive and lacks quantification. The current pre-demolition audit on the building basis is not efficient enough for large-scale mapping. As such, novel approaches for pattern identification need to be developed to facilitate contamination risk assessment in existing buildings. Data mining and its subfield machine learning present a new opportunity for using detection records to screen the likely presence of in situ hazardous materials in the national building stock.

The aim of the study is, therefore, to explore the potential of applied machine learning for predicting hazardous materials using building registers as input data and hazardous waste inventories as training and validation data. Considerable efforts have been dedicated to reducing the data uncertainty in merging and matching empirical data and building registers. The workflows of constructing a hazardous material dataset and a machine learning pipeline highlighted the complexity of processing unstructured, heterogeneous building-specific data. The results indicated that machine learning techniques succeed in characterizing suspected hazardous building materials, which is of significance for realizing the EU Construction and Demolition Waste Management Protocol. The detection likelihood of asbestos, PCB, CFC, and mercury were estimated according to inventory document types and building classes. Considering the building stock's diversity, a cross-validation matrix evaluating the quality and quantity of data subgroups was created for data stratification. Asbestos and PCB-containing materials in multifamily houses, schools, and commercial buildings were potential for modeling. Six supervised algorithms were used to test the prediction possibility. The average validation accuracies are 74% and 83% for predicting asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings.

Influential features to the prediction results were also visualized for expert knowledge interpretation, which has a practical implementation for assisting decision-making in constructing clean material loops. Construction year, floor area, and the number of stairwells and floors were influential for asbestos pipe insulation prediction, while construction year, balanced ventilation system, floor area are critical for PCB joints or sealants prediction. The proposed bottom-up modeling approach can potentially be scaled up to a national level and replicated in other countries with similar access to building stock data.

List of Publications

This licentiate dissertation is based on the following papers, referred to by their roman numerals in the text. The papers are appended at the end of the dissertation.

I *Machine Learning in Hazardous Building Material Management: Research Status and Applications*
P-Y. Wu, K. Mjörnell, C. Sandels, and M. Mangold
Recent Progress in Materials 3 (2), 24 (2021)

II *A Data-Driven Approach to Assess the Risk of Encountering Hazardous Materials in the Building Stock Based on Environmental Inventories*
P-Y. Wu, K. Mjörnell, M. Mangold, C. Sandels, and T. Johansson
Sustainability 13 (14), 7836 (2021)

III *Predicting the Presence of Hazardous Materials in Buildings using Machine Learning*
P-Y. Wu, C. Sandels, K. Mjörnell, M. Mangold, and T. Johansson
Submitted to *Building and Environment* (2022)

Other related publications by the author:

Tracing Hazardous Materials in Registered Records: A Case Study of Demolished and Renovated Buildings in Gothenburg
P-Y. Wu, K. Mjörnell, M. Mangold, C. Sandels, and T. Johansson
J. Phys.:Conf. Ser. 2069 (012234) (2021)

Acknowledgments

This licentiate dissertation investigates the interdisciplinary field of hazardous materials in the building stock with data mining approaches. The subject is niche but risky, in which hypothesis-driven theory and data-driven reality take turn navigating the exploratory process. Many people have contributed to making the research journey fruitful and inspiring.

I would like to thank my three supervisors for dedicating their expertise to the "patchwork". Kristina Mjörnell imparts the domain knowledge and demonstrates an extraordinary research paradigm. Mikael Mangold incubates the research idea and guides me through a broad spectrum of building stock analysis. Claes Sandels shares experience in applied machine learning and raises critical questions to orient the research direction. In addition, I am grateful for Tim Johansson's great work in merging building registers. Without him, the national building database won't be available for machine learning modeling.

I would also like to acknowledge the affiliation support from Lund University and RISE Research Institutes of Sweden for nurturing a young researcher. I am grateful for their collaborative efforts in providing a motivating learning environment. Besides, I am beyond thankful for Gothenburg City Archive, Stockholm City Archive, and the summer intern Frida Palstam for assisting data collection. Your endeavors are valuable for paving the scientific frontline forward.

Professional development cannot be progressed without the support from the dearest ones. I am sincerely grateful to my family and my babysitters for their unconditional respect and support for my choices to live abroad. We overcome the spatial distances and time difference, and bring our hearts closer. Also, my energy-booster, Amir, you enlighten my academic pursuit and our everyday life. We share passions and humor for research. Love from you all gives me hope, joy, and strength.

Pei-Yu Wu

Gothenburg, January 14th, 2022

This research was funded by the Swedish Foundation for Strategic Research (SSF), grant number FID18-0021.

Definitions and acronyms

Swedish real estate taxation register (Real property register)

The Swedish real estate taxation register includes information on tax data transferred from the Swedish Tax Agency to the Swedish Cadastral and Land Registration Authority.

Municipal cadastral register (Property map)

The Municipal cadastral register was reported from municipalities to the Swedish Cadastral and Land Registration Authority for the property map data product updates.

Semi-selective demolition

Semi-selective demolition is when demolition companies selectively collect all hazardous substances and that part of the non-hazardous substances that would overly reduce the quality of the stony fraction.

Backfilling

Backfilling is a recovery operation where waste is used as a substitute for non-waste materials to reclaim excavated areas or for engineering purposes in landscaping.

Artificial neural network (ANN)

ANN is derived from biological neural networks that have neurons interconnected in various layers of the networks. It can be used for both supervised and unsupervised learning.

Deep neural network (DNN)

DNN is a class of ANN algorithms for complicated learning tasks that simulates human neurons and forms the networks of multiple input layers, hidden layers, and output layers.

Multilayer perceptron (MLP)

Multilayer perceptron is a class of feed-forward neural network for supervised learning that consisting of an input layer, an output layer and one or several hidden layers.

Convolutional neural network (CNN)

Convolutional neural network is a class of artificial neural networks designed for processing structured arrays of data such as images.

Support vector machine (SVM)

SVM is a supervised learning classifier that projects the data points in space and determines their categories based on the gap for regression or classification.

Recursive feature elimination (RFE)

RFE is a feature selection method that fits a model and removes the weakest features until the specified number of features is reached.

Extremely Randomized Trees Classifier (Extra Trees)

Extra Trees is a tree-ensembled machine learning algorithm that combines the predictions from many decision trees fitted on the entire training dataset.

K-Nearest Neighbors (k-NN)

k-NN is a non-parametric supervised learning classifier estimating the likelihood of regression and classification based on what group the data points nearest to it belong to.

Extreme Gradient Boosting (XGBoost)

XGBoost is a tree-ensembled algorithm optimizing regularized gradient boosting for regression and classification tasks.

Receiver Operating Characteristic curve (ROC curve)

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier with varied discrimination threshold where the true positive rate at the y-axis is plotted against the false positive rate at the x-axis.

Area under the ROC Curve (AUC)

AUC is a scale variable estimating the overall performance of a binary classifier by representing the degree or measure of separability with a range between 0,5-1,0.

Shapley Addictive exPlanations (SHAP)

SHAP is a framework that explains the output of machine learning models using Shapley values, a game-theoretic approach used for optimal credit allocation.

Partial Least-Square-Discrimination Analysis (PLS-DA)

PLS-DA is a dimension reduction technique used for classifying categorical dependent variables.

Soft Independent Modeling of Class Analogies (SIMCA)

SIMCA is a statistical method for supervised classification for data with a set of attributes and their class membership for data labeling.

Linear/Quadratic discriminant function analysis (LDFA/QDFA)

LDFA is a classification and dimensionality reduction technique and QDFA is a variant of LDFA that allows for non-linear separation of data.

Random Forest (RF)

Random Forest is a supervised tree-ensembled algorithm that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.

Naïve Bayes (NB)

Naïve Bayes is a probabilistic supervised learning algorithm based on Bayes theorem for solving classification problems.

Principal component analysis (PCA)

Principal component analysis is a dimensionality reduction unsupervised learning method used for reducing the dimensionality of datasets by transforming a large set of variables into a smaller one without losing much of the information.

Accuracy (ACC)

Accuracy is measured by the number of true positives and true negatives divided by the total number of data points in a dataset.

$ACC = (TP + TN) / (P + N)$

TP: True positive
TN: True negative
P: Positive
N: Negative

Recall (REC, Sensitivity)

Recall is measured by the number of true positives divided by the total number of actual positives.

$REC = TP / (TP + FN)$

Precision (PRE, Positive predictive value)

Precision is measured by the number of true positives divided by the total number of positive predictions.

REC = TP / (TP + FP)

F1 score

F1 score is a harmonized mean of precision and recall and works well for class imbalanced data. F1 score is a scale variable with a range between 0,0-1,0.

F1 = 2 (REC * PRE / (REC + PRE))

Pseudo-R^2

Pseudo-R^2 is a performance measure for logistic regression based on the log-likelihood for the model compared to the log-likelihood for a baseline model using the formula:

pseudo R^2 = 1-(MSE/Var(Y))

MSE: average square error

Var: variance

Y: a set of variables.

CE	Circular economy
EPC	Energy Performance Certificates
PCB	Polychlorinated biphenyls
CFC	Chlorofluorocarbon
ACM	Asbestos-containing material
CDW	Construction and demolition waste
ML	Machine learning
BIM	Building information modeling
GIS	Geographic information system
HVAC	Heating, ventilation, and air conditioning
PV	Photovoltaic
AI	Artificial intelligence
PRISMA	Preferred reporting items for systematic reviews and meta-analyses

1. Introduction

In Sweden, a majority of buildings were built between 1945-1980 [1]. In the existing building stock of around 641 million square meters, the detached houses account for the largest share (41%). The multifamily houses have the second-largest share (33%), while the rest of 26% belongs to premise buildings. Nearly three-quarters of the heated areas in the Swedish building stock are older than 40 years and were built before 1980. In fact, half of the multifamily building stock was built from 1941 to 1970, and around 60% of the premise buildings were built before 1981. During this construction peak, numerous building components containing hazardous substances were produced and mounted in buildings [2]. Estimating their likely presence in end-of-life building stock before entering the waste stream is an up-to-date issue to be addressed. To characterize their detection patterns in diverse building types, developing a tailored approach according to their building tectonic typologies for risk assessment may act as a starting point [3].

The goal of this work is to use building stock registers to make building and component-specific predictions of the presence of hazardous materials. This was done using a linked subset of hazardous waste inventories from buildings with known detection/non-detection records as training data in machine learning models.

The opportunity to quantify detection risk according to building classes lies in the homogeneous characteristics of the building stock in Sweden. The Swedish dwellings are highly standardized and mostly inherited from governmental housing initiatives in the last century [4], [5]. In spite of uniformity in architectural design, they cannot be represented with a single reference building or a specific building technique [4]. However, there exist patterns regarding construction and material used in the Swedish residential building stock, especially those built in the construction segment [6]. Systemized construction methods for exterior wall and roof construction were recognized in multifamily houses, i.e., lightweight concrete walls with rendered façade or concrete sandwich walls, and detached houses, i.e., insulated wooden walls with clay brick or wood façade [4]. Also, by pairing the construction year, type of building materials, number of floors and apartments, one can categorize the Swedish multifamily houses into several building types: small house, slab block, panel block, and tower block [7]. The use of specific construction methods and the choice of materials for a particular building type may relate to the presence of hazardous materials to various extents. Using data mining as a knowledge discovery tool, the hidden relationships between building parameters from large-sized and multi-attribute sources can be generated [8].

1.1. Construction and Demolition Waste Management

Considerable resources are used to construct, operate, and maintain buildings. By the end of their service life, building structure and components are, in most cases, abandoned or demolished. The conventional degrading handling process for construction and demolition waste (CDW), such as energy recovery and backfill disposal, leads to ineffective material value recovery. Therefore, a call for closing the linear loop has gained more attention in the construction sector [9]–[11]. The drivers behind the transition are a limited supply of raw materials as well as an increasing cost of waste handling [12]. New possibilities to facilitate urban metabolism and renewal are explored in a growing body of literature, including material stock and flow analysis in buildings [11], [13]. However, numerous barriers in cultural, regulative, financial, and sectoral aspects are required to be addressed to materialize a circular framework in the built environment [14]. These are, for example, underdeveloped CDW regulations, inconsistent data quality, incomplete reverse logistics, and low market readiness for secondary materials [15].

Nowadays, construction and demolition waste remain the largest waste stream, accounting for 30-40% of total solid waste worldwide [16]. Taking EU countries, for example, 36% of solid waste can be traced back to the construction industry, yet less than half (46%) of the CDW were recovered (including backfilling) in 2016 [11]. Also, a huge discrepancy in recycling rate was observed among the countries and an unrepresentative recycling rate by taking backfilling, where waste is used as a substitute for non-waste materials in construction, into account [17]. Even though CDW is a significant source of secondary material, the barriers of reuse and recycling remain mainly due to various definitions of CDW between countries, demanding coordination between actors, and disparities in data collection approaches [17]. Another significant impediment can be attributed to a lack of confidence in the quality of recycled CDW materials [18].

To address these challenges and reduce the environmental impact of current construction practice, the EU Waste Framework Directives and multiple related action plans were established [19], [20]. The EU Communication “Resource efficiency opportunities in the building sector” (COM 445, 2014) is regarded as the cornerstone for the sectorial transition, where the roadmap to achieving 70% of the recycling target in 2020 is created [21]. Under the legislative framework, the EU Construction and Demolition Waste Management Protocol [20] and the Guidelines for the Waste Audits before Demolition and Renovation Works of Buildings [18] were published to accelerate the implementation of associating measures along the waste value chain.

With the primary objectives for enhancing the market adoption of secondary materials and reliability of CDW management practice, five interlinking actions are formulated in the EU Construction and Demolition Waste Management Protocol: (1) improved waste identification, source separation, and collection; (2) improved waste logistics; (3) improved waste processing; (4) quality management; (5) appropriate policy and framework conditions. The sequential steps correspond to the general flow of CDW processing and require comprehensive input from stakeholders. To exploit the potential of recycling and reuse of CDW, the first action plays a crucial role in ensuring the quality of the waste [2]. Identifying and removing the undesired waste fractions help position the trajectory for a functional circular practice.

1.2. Building Stock Information

Over time, the city composition changes constantly to adapt the society's development. On the one hand, the building stock study represents an integral subject where multiple disciplines are intertwined; on the other hand, it can be viewed as a parchment where layers of urban metabolism were capsulized along the buildings' lifecycle. Demolition becomes an inevitable option to adapt to urban growth challenges, including building functionality improvement, demographic structure change, and hazard decontamination [22]. Mining the values from clean end-of-life buildings meets the trend of circular transition [23]. From the perspective of resilient building stock management, the existing building stock can be considered as a secondary resource [24]. Hazardous material inventories is therefore a prerequisite to facilitate abatement measures for the existing building stock [25]. For instance, Donovan and Pickin [26] demonstrated the metabolism of the asbestos stock through material flow modeling on a national scale, and Diamond et al. [27] addressed the policy measures by estimating the source of PCB in the citywide building stock.

Furthermore, conducting data mining on existing building registers promotes building stock analysis. Some registered data are digitalized and accessible in Sweden owing to the long tradition of documentation. The registered data are preserved by central authorities such as Statistics Sweden, Swedish Cadastral and Land Registration Authority, Swedish Tax Agency, Swedish National Board of Housing, Building, and Planning, and the old building documents are stored in municipality archives and municipality museums. In the past decade, the extensive implementation of Energy Performance Certificates (EPC) has improved the data granularity and facilitated data comparability to a great extent in the dimensions of building components. Previous studies have validated the Swedish EPC data and minimized data uncertainty [28]. Fostered by data analytics and the need for evidence-based policy instruments, the applications

of EPC have exceeded its original intention [29]. Through adjoining information of hazardous materials and components from the pre-demolition audit inventories to the building database, the understanding of the presence of such materials in the existing building stock can be enriched. By pairing the municipal cadastral register, the real estate taxation register, the EPC data, and the inventories of hazardous waste, new approaches for material risk assessment in the building stock can be developed to enhance material recyclability [25]. Also, the data-driven assessment methods can complement the limited adoption of the present environmental investigation by mapping potentially contaminated buildings.

Several impediments are required to be overcome to succeed with data mining in the construction sector. Among all, poor data quality of construction datasets was highlighted in terms of missing or misleading values [30] and insufficient validation [29]. Also, the data uncertainties can be traced back to errors in data collection or documentation [28], as well as matching between multiple heterogeneous data sources [31]. Improving the data reliability and consistency of existing building require massive efforts, yet it lays a critical foundation for building stock modeling [25]. Compared to the data quality aspect, other essential aspects, such as knowledge interpretation and method generalizability, are given insufficient attention [8]. Domain knowledge acts as a means for explaining factorial correlations and bridging the gap between data-derived rules, which guide the data mining process and transform the data insights into scientific outputs. In data embedded case studies, having an overarching picture on data representativeness, replication possibility, and result applicability can improve model interoperability in other contexts. This will benefit the validation of the results between comparative studies and thus stimulate a good data mining practice. In short, the elements of data quality, domain knowledge incorporation, and case representativeness underpin the success of machine learning modeling.

1.3. Hazardous Materials in the Building Stock

Over the years, the presence of hazardous materials has posed significant challenges and concerns for deconstruction, renovation, and demolition projects [2]. The unforeseen encountering of hazardous materials during renovation or demolition requires acute decontamination, which leads to unexpected project delays and could account for as much as 20% of the demolition cost [32]. The fundamental problems are attributed to the opaque content description of the building products [33], along with the complicated waste sorting process [34]. Digitalized building stock with the use of Building Information Models (BIM) for end-of-lifecycle scenarios becomes an alternative direction for minimizing

the CDW in the future [35]. The creation of the intelligent and object-oriented models is based on building information modeling methodology, where digital 3D models containing geometric information and non-geometric properties of all the building elements can be created [36]. Reliable and error-free data from BIM are considered as potential sources to simulate circular materials from end-of-lifecycle buildings [35].

Despite of emergent evaluation of tools and inspection protocols for the quantification or measurement of materials, their applications remain restricted for individual buildings. In short of a global framework, the introduction of material passport [37] and component bank (BAMB) [38], as well as BIM-based end-of lifecycle prototypes, have currently limited application at local and case-specific conditions [35]. Also, the disconnection between BIM and end-of-lifecycle management tools has been reported as a hurdle to their use in the existing building stock [39]. Alternatively, exploiting material inventories from buildings is rather vital for paving the circular ambition forward. Developing appropriate risk management tools to estimate the uncertainty of reusing or recycling contaminated materials for the newly-formed circular economy chains is necessary and urgent [40].

To overcome the limitations of BIM-based end-of-lifecycle deconstruction applications and further incorporate risk evaluation for existing buildings, pre-demolition audit is viewed as a viable alternative. Pre-demolition audit (or waste audit) refers to an environmental investigation where hazardous substances and materials are assessed prior to renovation or demolition [2]. The pre-demolition audit process usually consists of a desk study on original building documentation and maintenance protocols, a field survey for inventories of hazardous waste with potential sampling and analysis. Then based on material assessments and quantity estimation, the auditors provide management recommendations and reporting [2], [25]. The generated inventories of hazardous waste are not only used for material recovery, but also for planning safe deconstruction works. Complete documentation of hazardous waste inventories should therefore contain the following information: (1) report on the suspect and identified hazardous components concerning their amount and location; (2) report on potentially reusable and recyclable materials along with estimated treatment costs; (3) market research on different options for waste management [41]. In practice, the pre-demolition audit has been performed mandatorily or voluntarily in several European countries. In Sweden, the inventory of hazardous waste is obligatory with guidance on the implementation scope, worker safety, search list for hazardous materials, and sampling approaches [42]. Given its substantial benefits for the CDW management, the ultimate goal is to develop a harmonized pre-demolition protocol for cross-regional waste auditing, as well as an international auditor certification system for contractor evaluation [2].

Hazardous materials require special care both during onsite selective demolition and offsite waste sorting. The risk of secondary contamination can be minimized if building components with the primary contaminants are removed intact before deconstruction. Therefore, regulations for safe management and disposal of certain hazardous materials, i.e., asbestos and PCB-containing materials or lead-based paint, [43] are imposed. Yet, disconnected legal frameworks between CE initiatives and the EU REACH regulation (Registration, Evaluation, Authorization, and Restriction of Chemicals), along with the complexity of risk management between involving actors, were substantial barriers highlighted in the literature [40].

The frequent use of hazardous materials in the past considerably increases the risk of contaminating exposure in the existing building stock. Table 1.1 shows an overview concerning the use and ban of particular hazardous materials in Sweden [44]. The current environmental investigations undertaken in individual buildings for the inventories of hazardous waste fail to address large-scaled predictive maintenance planning. Considering the retrofit needs for the aging post-war buildings and the corresponding growing hazardous waste stream, developing an efficient way to screen remaining in situ hazardous material prior to deconstruction is favorable. As a result, researchers worldwide are searching for new approaches to quantify the extent of the contaminated building stock [32], [45]–[47]. Acquiring this kind of hazardous material information can add substantial values for decision-making in both policy implementation, i.e., decontamination and abatement, and practical utilization, i.e., deconstruction or semi-demolition. With the ambition of improving the recycling rate of CDW under the EU policy framework, it is beneficial for authorities to organize initiatives for hazardous building stock decontamination. On the other hand, property owners require the information for risk assessment in the demolition or renovation permit application. Understanding the approximate location and amount of hazardous materials and components in buildings allows the estimated project schedule and cost to be better controlled.

Table 1.1 An overview of the use and the ban of asbestos, PCB, CFC, and mercury-containing materials in Sweden [44].

	1920	**1930**	**1940**	**1950**	**1960**	**1970**	**1980**	**1990**
Asbestos	*1976: Ban of crocidolite, 1986: Total ban of all asbestos products*							
Pipe insulation	1920s						1986	
Cement panel		1930s					1986	
Tile/clinker	1920s					1976		
Carpet glue	1920s					1976		
Floor mat	1920s					1976		
Ventilation channel	1920s					1976		
PCB	*1972: Ban of PCB, 1978: Last use of PCB in electronic equipments*							
Joint/sealant				1956		1975		
Double glazing windows				1956		1977		

Capacitors	1920s				1995
Acrylic flooring		1956		1975	
CFC	*1990s: Ban of CFC*				
Fridge/freezer			1960s		1995
Building insulation			1960s	1970s	
Cooling unit			1960s		1995
Mercury	*1993: Ban of mercury*				
Relay/switch	1920s			1970s	
Pressure gauge	1920s				1993

1.4. Inventory of Hazardous Materials in the Swedish Building Stock

The concern of residual hazardous materials has long existed in Sweden. The major Swedish residential building stock, especially multifamily houses, inherits from the post-war era 1945-1960 and 1964-1975. These construction periods collide with the massive use of hazardous materials, i.e., asbestos, PCB, CFC (chlorofluorocarbon), and mercury. Due to their long lifespan and stable properties, these hazardous materials enter the waste stream or re-contaminate other building components after many years of deconstruction [25]. Over the years, they have caused many delays and high decontamination costs in renovating and demolishing buildings, as well as environmental and occupational health risks [48]. In view of the ongoing renovation wave of the existing building stock, along with the requirements for clean CDW for a circular economy, characterizing the presence of hazardous materials and taking decontamination measures in advance is rather urgent.

Since the mid-1990s, Sweden has introduced obligatory pre-demolition audits for renovation and demolition building permit applications. The fundamental legislations for CWD in Sweden trace back to the Building Code (Miljöbalken, SFS 2010:900), the Planning and Building Act (PBL), and the Waste Ordinance (Avfallsförordningen). Together with the requirements for a safe workplace during demolition and rebuilding from the Work Environment Acts (Arbetsmiljölagen), the Swedish Circulation Council in Construction (Byggsektorns Kretsloppsrådet) published practical guidelines to assist resource and waste management for construction and demolition projects [42]. The pre-demolition auditing practice has resulted in considerable inventory data with a high potential for hazardous material assessment at the stock level. Nevertheless, these data remain difficult to access and unexplored as the environmental information is not digitalized nor connected to national building registers. The hardcopy or scanning pre-demolition audit documents are stored individually in each municipality archive. Compiling these hazardous waste inventories and

validating their quality is an enormous task, yet the work can contribute substantial value to in situ hazardous material management.

The inventory of hazardous waste intends to pinpoint which materials and components contain hazardous substances that require special care. Executing a compulsory environmental investigation applies to all buildings, and the results should be appended to demolition or control plans. The materials that are considered human or environmentally hazardous waste can be referred to the waste ordinance (Avfallsförordning, 2011:927) [2]. Extremely harmful substances, such as asbestos (AFS 2006:1) and PCB (polychlorinated biphenyls) (SFS 2007:19), have additional ordinances concerning decontamination measures and waste disposal. The progress of CDW management is seen in certification systems for the current building stock and qualification of auditors [2]. For instance, the Sweden Green Building Council launched the Environmental Building Operation and Administration Certification for the existing buildings (Miljöbyggnad iDrift) to evaluate the environmental impact caused by materials and introduce measures to reduce waste generation [49]. An inventory of hazardous waste and a waste management plan is required for all buildings to maximize the waste recycling potential, yet following the industrial agreement on the Swedish Circulation Council in Construction's guidelines (Kretsloppsrådets riktlinjer) during the environmental investigation is voluntary [2].

1.5. Research Focus

Hazardous material identification in the existing building stock is crucial for realizing circular construction [10]. Abatement measures for hazardous material before deconstruction can considerably reduce the risk of secondary contamination. Therefore, the first and foremost action in the EU Construction and Demolition Waste Management Protocol [20] underlines the importance of waste identification, source separation, and collection. The work in the licentiate thesis is dedicated to method development for assessing the risk of in situ hazardous materials in buildings. Three intercorrelated research opportunities were explored in the study:

The first research opportunity relates to the data availability from hazardous waste inventories. The growing recognition of developing quality-assured CDW management advances the legislation at the EU level, meanwhile gradually facilitating a paradigm shift in the construction sector. A new possibility emerges for data-enabled building material stock investigation and mediation. Creating a dataset from the past detection records allows probing the usability of the hazardous waste inventories for estimating the potential presence of the remaining hazardous materials in the building stock. Screening

emergent applications and relevant data in the CDW management field are addressed in Paper I and Paper II.

The second research opportunity concerns the possibility of assembling a national building database with information on hazardous materials in buildings. The methodology for merging multiple building registers and the associating challenges have been developed and discussed in the literature [31], making the coupling between the general information and specific information viable. The acquired building registers can be used to construct a prediction dataset for evaluating the risk of hazardous materials on a national scale. A workflow for using the data from inventories of hazardous waste is proposed in Paper II.

The third research opportunity pertains to the progress of artificial intelligence and computational power. Machine learning algorithms show a promising capability to predict unknown examples based on the labels from past data. With the help of supervised models, the complex causality between predictive variables and target variables can be untangled. The previous detection records from inventories of hazardous waste are used to draw insights into influential factors and characterize the likely contaminated buildings. Machine learning model development and evaluation are addressed in Paper III. Exploiting the three research opportunities enables identifying the potential occurrence patterns of hazardous materials. Paper I offers insights on innovative data-driven approaches for hazardous material recognition and management. Paper II bridges the gaps between CDW management and building stock analysis by adding data from inventory of hazardous waste to the generic building stock information. Paper III tests the feasibility of using ML to construct a prediction pipeline. The research outcomes will help the property owners and demolition companies to plan abatement measures without exposing them to the risk of project disruption or the health of workers. A conceptual diagram showing how the study fits into the overarching trends is illustrated in Figure 1.1.

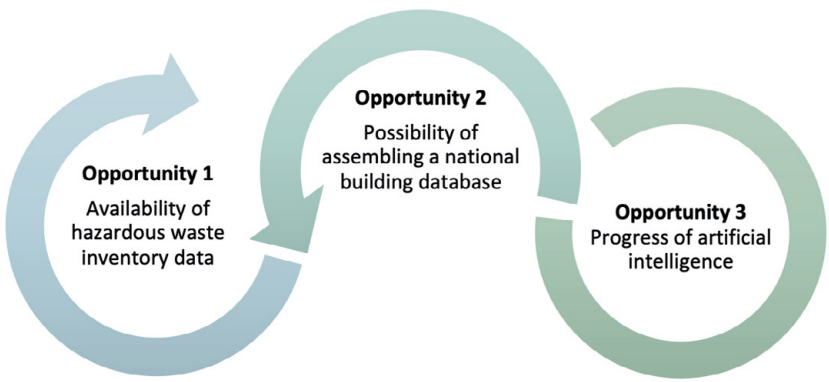


Figure 1.1 The three research opportunities exploited for investigating the presence of hazardous materials in the building stock.

1.6. Aim and Research Questions

Aim The thesis aims to understand occurrence patterns of hazardous materials in the Swedish building stock using machine learning techniques to deliver decision support to the relevant actors managing material circularity.

The first research question sets a theoretical background for methodology screening in the field. The second and the third research questions elaborate on how data mining contributes to the machine learning pipeline development.

RQ1 What are state-of-the-art data-driven applications for hazardous material management?

RQ2 What is the potential to use data from hazardous waste inventories to assess the risk of hazardous materials in the building stock?

RQ3 How accurate can asbestos and PCB-containing materials in specific building classes be predicted using machine learning models?

Figure 1.2 illustrates the association between the research questions, the appended papers, and the major themes – method screening, dataset creation, model development, and attempted prediction.

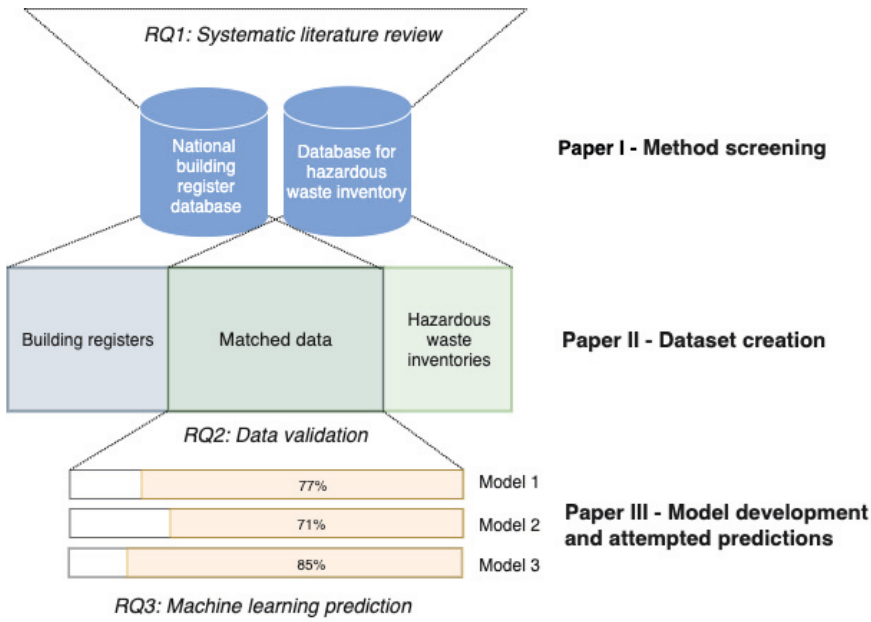


Figure 1.2 The research process describes how the research questions entangle with the major themes and achieve the study goal.

1.7. Challenges and Limitations

This PhD project is deeply explorative and hypothesis-driven. It is not certain that the patterns can be identified and used to predict the presence of hazardous materials. In addition, there are several challenges and limitations concerning data representativeness, uncertainty, and heterogeneity that need to be mitigated:

- Data representativeness

The primary limitation of the study involves data representativeness. The hazardous waste inventory data were obtained from demolished or renovated buildings; thus, the detection records could not be fully representative of the entire building stock. To prevent the risk of false inference in the upscaling process, the metadata between samples and large building stock was controlled.

- Data uncertainty

The secondary concern relates to data uncertainty. Since no standard protocol exists for pre-demolition environmental auditing, the inventory types for different building classes and the competence and experience of auditors vary. The heterogeneous assessments from individual buildings over time challenge the data quality and completeness. The effects of data quality and quantity on the prediction results are visualized and minimized through matrix validation.

- Data heterogeneity

However, one should be aware that building stock study is, by its nature, context-dependent. The prediction results may only apply to the Swedish building stock, yet the developed machine learning approach is universal and can be replicated to analyze the presence of hazardous material in other building stocks. Therefore, this thesis focuses on the challenges and how they were overcome in the method development.

1.8. Content Structure

In the introductory chapter, the background information about building stock information, hazardous waste inventories, and the research needs for identifying in situ hazardous materials were illustrated. In Chapter 2, the previous research regarding data-driven approaches and their applications in CDW management were reviewed. Chapter 3 describes the methods for data acquisition, compilation and validation, and their use in the machine learning pipeline. After that, the key findings were summarized in Chapter 4 and discussed in Chapter 5, where contributions and practical implementation were highlighted. Lastly, in Chapter 6, the conclusion and in Chapter 7, suggestions for future research were presented.

2. Previous Research

In this chapter, a review of the former studies that shape the research landscape of the interdisciplinary field is presented. Two major subjects – pre-demolition audit in construction and demolition waste management (Section 2.1) and data-driven building stock analysis (Section 2.2) – form the basis of the research scope of pattern identification for hazardous materials occurrence. The research gaps of hazardous material recognition and the practical challenges in recycled material quality assurance were highlighted and addressed by exemplifying applied machine learning methods and their associated data at the research fronts.

2.1. Pre-demolition Audit in CDW Management

The divergent barriers along the value chain of the construction and demolition waste management hinder material recovery and recycling [15]. The foremost issue concerns the high cost and time-consuming process for mixed waste sorting [34]. Consequently, semi-selective demolition – removing hazardous substances and part of the non-hazardous substances that could reduce the material fraction's quality – becomes a viable alternative to ensure secondary material purity and traceability [12]. With the gradual adoptions of economic incentives and legislative obligations, the concept of in situ material management was proposed to replace conventional total demolition and offsite sorting. Uptaking this trend, researchers develop prediction models and assessment methods for waste generation and hazardous material identification. The following subsections are structured as CDW estimation in the circular economy framework (Section 2.1.1), applications of AI in Hazardous Materials Management (Section 2.1.2), as well as approaches to hazardous material identification (Section 2.1.3).

2.1.1. CDW Estimation in the Circular Economy Framework

The importance of pre-demolition audit to ensure a functional circular economy framework is particularly emphasized and acknowledged by CDW practitioners. Nevertheless, the manual process of material volume measurement and document retrieval in pre-demolition audits demands a

lot of time and effort. Moreover, demolition and renovation waste is usually estimated for specific projects that may hamper the results usability and generalization in broader applications [50]. The research gap in the CDW estimation methods lies in low application to the existing building stock. Often, data at the building aggregation level were lacking for old buildings, not to mention the fact of the low availability of their BIM models [35], [51]. Therefore, screening the empirical data from the past demolition projects and validating data quality may be an alternative source of information for waste analytics [50], [52].

The efforts to investigate CDW data accessibility and reliability have been put in place in earlier research. Sáez and Osmani [15] examined the Eurostat CDW data quality and analyzed the CDW generation and recovery rates. The results were then used to assess the performance of selected EU countries against the EU CDW recovery target and countrywide CDW policy frameworks. The concern over the plausible CDW data has been explicitly shown in terms of low data quality and harmonization. Likewise, the prevalent problems of missing data and information barriers hinder the decision-making for construction waste management [53]. To address the issue of inevitable information gaps, Yang et al. [30] applied behavior-based machine learning methods to process project-level structural missing data from aggregated waste generation behaviors. Automatic feature selection was applied to extract key waste generation behavioral features and succeeded in missing value prediction (*F1 score* = 0.8-0.87) [30]. These studies demonstrate the potential applications of waste data, meanwhile pinpointing the immature development of a standardized CDW data protocol. The essential role of CDW data in handling recycled waste should not be overlooked.

Despite the current data limitations, new attempts have been made in predicting CDW generation from end-of-life buildings. Obtaining the estimated amount of waste and type of information can facilitate the short time constraint for building material removal and recovery [50]. Supervised learning algorithms and deep learning algorithms advance the prediction of the amounts of recovered materials before demolition. Cha et al. [54] developed a prediction model for handling a small dataset with mixed data types. With few input features for different material types, the patterns for predicted and observed values can be recognized. The developed random forest (RF) models were of practical use when available data was limited, yet the robustness of the RF models needed to be verified with a larger data size [54]. On the contrary, the deep neural network (DNN) was tested by Akanbi et al. [50] for a similar study objective. By employing basic features of buildings to the DNN models for recyclable, reusable, and landfill waste materials, high-performance accuracies were achieved through evaluating a case study from a building given four archetypes – concrete, masonry, steel,

and timber [50]. In summary, the state-of-the-art predictive models contribute to data gap filling and support decision-making during the hazardous waste inventory practice.

2.1.2. Applications of AI in Hazardous Materials Management

Holistic approaches were explored to address the integration of sustainability and safety aspects for effective hazardous material management. To address the risk of chemicals retained in the material cycles or re-entered in the environment, Bodar et al. [40] proposed a safety decision scheme to enable risk evaluation of chemical products and wastes in the circular loop for stakeholders. The bridge between EU legal frameworks, such as Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) and Substance of Very High Concern (SVHCs), and their interfaces with circular economy required to be constructed [40]. Furthermore, a disaster-response program for asbestos-containing material (ACM) management was established by Kim and Hong [55] to enable practical information transmission. Through a six-fold case study, the location and dismantle priority of the buildings with ACM, the types and quantities of ACM, as well as the asbestos fiber and greenhouse gas emission for ACM removal, were investigated [55]. From these examples, a shift of hazardous material management from traditional contaminant monitoring and abatement to precautionary principle-embedded risk management is observed.

To progressively increase material recycling rates while keeping the material loop clean requires a synergy between building stock management and an adapted demolition process. The adoptions of BIM and Geographic Information System (GIS) allow stock material information retrieval for the single building or the built environment in the city as a whole [35]. Hence, Rašković et al. [25] evaluated data collection methods to obtain input data for as-built building information modeling. Their findings showed that the demolition-related information could be extrapolated and used to enrich the existing building information databases by incorporating the building material assessment methods and geometric data capture tools. A case study in a test building was presented to illustrate the proposed workflow of merging geometric information and pre-demolition audit information for 3D modeling. A bottom-up approach of data collection and model development for the existing building stock provides more detailed characteristics; thus can be more suitable for deconstruction and selected demolition in practice [25].

2.1.3. Approches to Hazardous Material Identification

According to the EU Construction and Demolition Waste Management Protocol, quantifying in situ hazardous materials consists of– material identification, source separation, and onsite waste collection. The previous studies regarding hazardous material identification can be approached from top-down and bottom-up perspectives. Remote sensing within the geomatics and information field allows efficiently quantifying suspect hazardous materials from aerial, hyperspectral, or multispectral images [56]. The method for recognizing asbestos-cement roofing using multispectral imaging and field inventory was validated by Fiumi et al. [56]. Around 89.1% of general accuracy was attained for material classification in the cross-comparison study. Comfirming the method's potential, a similar study was conducted by Wilk et al. [57] to identify the critical factors associated with the amount of asbestos-cement roofing, laying a foundation for asbestos stock estimation in Poland. Regional asbestos-containing cement roofing was mapped out by matching the aerial or satellite images and the field inventories [58]. The number of individual farms in the village, the distance to the asbestos manufactural plants, the building age, as well as local social-economic situation affect the use of asbestos-cement [57], [58]. Strong spatial clustering was found between malignant mesothelioma, a deadly tumor cancer caused by asbestos exposure, and the location of asbestos manufacturing plants in the geostatistical analysis [59]. By employing these features in the random forest models, the prediction map of the spatial distribution of asbestos-containing materials (ACMs) at the national level can be constructed [58]. Further on, Krówczyńska et al. [60] tested deep learning algorithms, i.e., convolutional neural network, for aerial photographs classification of asbestos-roofing and achieved comparable accuracy rates.

However, remote sensing is limited in identifying broad assortments of hazardous material since many of them are visually unrecognizable [43]. Intrusive sampling and lab analysis are rather common ways when experience-based material diagnoses are not applicable. The detection records from pre-demolition audit inventories and the hazardous material description databases become valuable sources for data-driven studies. Through accumulating various input data from individual buildings in the citywide demolition database [32] and a questionnaire in a mobile application for assessing suspected asbestos materials [45]–[47], the likely contaminated building components in the residential environment can be pinpointed. Statistical methods and ontology-based approaches have also been adopted for this purpose. The previous empirical studies showed the prevalent occurrence (82.5-95%) of asbestos-containing materials in residential buildings [32], [45]. However, the detection frequency of specific asbestos components varied substantially from country to country [32], [45]. The identified asbestos type is primarily nonfriable chrysotile

[32] with a low potential for disturbance and removal priority [45] for residential buildings. Characterizing the extent of in situ hazardous materials can facilitate decontamination work planning before deconstruction or demolition. Other than using sampling approaches, Mecharnia et al. [61] developed an inference approach for predicting the probability of asbestos-containing materials based on temporal descriptions of the marketed products. Evaluating the method on actual data showed promising prediction results concerning the presence of asbestos suspecting products, locations, and structures in buildings.

With respect to source separation and onsite waste collection, the detection methods for optical identification of hazardous material and mineral images were established. An unsupervised learning algorithm, i.e., principal component analysis, was combined with hierarchical clustering analysis to separate asbestos-containing materials from the rest of CDW [62], [63]. A hybrid of supervised learning algorithms, i.e., support vector machine, and deep learning algorithms, i.e., convolutional neural network and multi-perceptron, also proved to be effective for CDW image detection [64]–[66]. However, low prediction rates were obtained in predicting recycled aggregates due to object variabilities, as shown by Anding et al. [67], [68]. Overall, Kuritcyn et al. [66] confirmed that the non-invasive image processing methods improve hazardous material recognition in selective demolition, meanwhile enhancing the safety in the recycling process of CDW.

2.2. Data-driven Building Stock Analysis

Technological convergence provides a new perspective for building retrofit and demolition, which are the constant activities taking place under legislative requirements and demographic dynamics. Emergent building stock data and the use of computation tools extend the scope of building stock analysis to an unprecedented degree [69]. Various information collected along the building lifecycle, from design, construction, commissioning, operation, and maintenance to retrofit, can be interlinked to enrich the building database. This section reviews the former research on building stock data screening and validation (Section 2.2.1), then highlights machine learning applications in the building stock analysis (Section 2.2.2).

2.2.1. Building Stock Information Screening and Validation

Data mining applied in the construction industry has led to an exponential growth of knowledge development in recent decades [8], [53]. Built upon the techniques of statistics, machine learning, and pattern recognition, it enables rules, correlations, relationships, and anomalies detection from unstructured and complicated data sources [70]. One of the principal sources, Energy Performance Certificates (EPC), underpin the EU building stock data owning to its comprehensive coverage, standardized scheme, and long-time span records [29]. Since its introduction in 2017 up to the launch of the second version, roughly 82% of buildings in Sweden are covered with the mandates [28]. Therefore, EPC has been used as a core data source to connect with other registered data, including building footprints [31], multifamily property and building data [31], [71], occupant socio-economic data [72], [73], and spatial information [56], [58]–[60], [74], to create a deeper understanding of varied thematic areas within building stock analysis. From the divergent lenses, scientific consistency can be controlled by evaluating the associations between identified patterns and domain knowledge [75]. However, the variant versions of measurements, aggregation levels, and updating frequency can put the matched data quality under question [53]. More efforts are required for harmonizing the differences to construct a trustable dataset.

Over the past decade, the diverse applications of EPC have expanded beyond its original intention to inform the actors in the building sector about building energy performance [29]. For example, the high potential was highlighted for using EPC data to untangle the causal relationship between building energy demand and other factors [76], as well as creating an overview and validating building stock models [28]. However, in the analyses of the performance gap between estimated and actual energy performance, the uncertainty and value discrepancy of EPC data were highlighted in the earlier research, specifically, heated area measurements [28], energy consumption, and energy conservation assessments [77].

Accordingly, Pasichnyi et al. [29] developed a data quality assurance method to fill in the gap of fundamental attributes to data accuracy and consistency. The proposed six validation levels refer to earlier efforts from Simon [78]: data structure check for limiting missing values, consistency check between values and records, comparative appraisal between dataset revisions, check between original data sources, check between domain, and lastly, check between data collectors using the shared keys. Their findings suggest that using auxiliary data can expose the underlying data problems, facilitating initial data quality control during data collection, data cleaning, and processing in data analyses. In relation to this, the EPC should be further refined to include the data quality parameter for its features [29].

Appending this metadata helps address data uncertainty due to defective data collection instruments, multiple sources, data entry errors, and avoid preprocessing time for data noise removal, suggested by Yan et al. [8]. General criteria on data reliability, completeness, consistency, and resolution are expected to enhance the overall data quality in the construction industry.

2.2.2. Machine Learning Approaches in the Building Stock Analysis

Machine learning, an application of artificial intelligence that features an automatic system learning from data and generates knowledge, offers a new perspective to building stock analyses [8], [69]. The applied machine learning approaches enable building stock analysis to overcome several barriers. Firstly, the capability to identify underlying relationships from small datasets. The fact of scarce data in the construction sector is a bottleneck to foster building stock analysis. The difficulty to access reliable data [8], resource-demanding data collection and processing, hard-to-recover missing data [30], and class imbalanced data [79] hinder the development of the field. Nonetheless, machine learning techniques enable pattern identification from historical records to predict future developments, making them promising tools for decision support and automation [80].

Next, various machine learning algorithms can process heterogeneous data types and amounts that those traditional statistical methods fail to do [8]. The learning settings duo regression and classification enable machine learning algorithms to untangle the complex relationships between numeral continuous and categorical variables, as well as text or image data [81]. The high flexibility of machine learning for input data results in extensive applications in most fields; building stock analysis is not an exception [69]. Leveraging the statistical learning from the subset of observations, the performance of predictive algorithms exceeds the rule-based control without the need to explicit the thorough assumptions [82], [83]. Finally, by linking various dimensions of building stock analysis with the input from environmental [84]–[86], economic [87], and social data [31], [72], [73] to machine learning models, a more holistic picture of the present, past, and future status of the building stock can be developed.

However, drafting a clear problem statement is a prerequisite for the creation of machine learning models. Understanding the strength and weakness of each algorithm is more likely to achieve optimal prediction results for the intended purpose. Generally, three learning methods are classified according to the learning problems and available data – supervised, unsupervised, and reinforcement learning [81]. Supervised learning, including support vector machine (SVM) and artificial neural

network (ANN), were found to be common methods for building simulation, diagnosis, and probability assessment when data labels are accessible [80]. The SVM and logistic regression classifiers were applied by von Platten et al. [6] to predict buildings features for energy efficiency strategies. In the study, data from the Swedish Land Survey, EPC and, auxiliary building observations from Google Street View were used as input to estimate specific building typology for multifamily houses. Similarly, classification purpose of prediction has been seen in a wide range of thematic areas, for instance, categorizing energy poverty risk based on socio-economic data by Longa et al. [72] and leveling the importance of the features for predicting building use and performance based on smart meter data by Miller [88].

Unsupervised learning for data transformation and clustering is often applied to building certification, such as building energy performance benchmarking based on building features shown by Gao et al. [89], or buildings' envelopes performance evaluation from infrared thermography field survey, climate and energy consumption data presented by Wang et al. [90]. Unsupervised algorithms are also seen to be combined with supervised learning algorithms for data preprocessing, particularly for data stratification and dimension reduction. Common techniques are for instance, principal component analysis (PCA), multiple correspondence analysis (MCA), k-means clustering, and k-medoids clustering. Kropat et al. [91] conducted a predictive mapping of indoor radon concentrations with k-medoids clustering for automatic classification, then used the random forests and the Bayesian additive regression trees for prediction.

On the other hand, reinforcement learning, featuring using agents to find optimal solutions in a predefined environment, is mainly employed in emergent monitoring or control studies [69]. Chen et al. [82] developed an optimal control decision model to regulate heating, ventilation, air conditioning (HVAC), and window systems to minimize energy consumption and thermal discomfort. The reinforcement control is proved to be more efficient than heuristic control and can immediately adapt to the changing environment. Finally, the deep neural network has been adopted extensively in many building stock thematic areas [80], including energy consumption [92] and regional energy market forecast [93], solar radiation [85] and photovoltaic (PV) power production [94], optimization of cost and CO_2 emission in the integrated energy-water consumption models [84], as well as demolition waste prediction [50]. It gains popularity for the ability to create new task-specific attributes from data representations [95], as well as the capacity to identify complex structures or relationships in high-dimensional data [50]. Leveraging the insights from these examples, the applied machine learning methods have been used to affirm analysis results from conventional approaches and further facilitate evidence-based

decision-making [6]. High flexibility in data aggregation level and operationality of time series data make machine learning a suitable method to study the existing building stock [69] since it is hard to characterize its features due to insufficient and incomplete building documentation [25].

The presence of hazardous materials in the built environment is one of the complex and longstanding problems. Even though the use of hazardous substances has been prohibited and regulated since the 1970s in many countries, the exposure risk remains in buildings' operation and post-use phases due to insufficient understanding of hazardous materials occurrence patterns [25]. In regard to this, short-term remediation strategies, such as air concentration monitoring [96], and long-term decontamination measures, for instance, semi-selective demolition [12], become a conventional practice to control the risk of in situ hazardous materials. However, with the need for improving the purity of recycled materials in a circular economy, a more comprehensive and cost-efficient approach is required to characterize the detection of hazardous materials [15]. Recent studies that trained supervised and deep learning models on remote sensing data accomplished large-scaled asbestos-cement roofing screening and estimation [58], [60]. Yet, the studies of asbestos-containing materials at the building level remain only on statistical quantification [32], [45].

The research gap for using machine learning algorithms to locate the possible detection of hazardous materials in the building stock is identified [61]. Another research area, namely employing EPC data and other building-relevant registers for the prediction of hazardous materials, is also unexplored. The possibility of applying machine learning methods and the potential input data requires to be further investigated. By using data mining on the past detection records for the binary classification, the risk of encountering hazardous materials in the entire, not yet inventoried building stock can be evaluated.

3. Research Methodology

The study design of the sequential work on applied machine learning method development with the focus on in situ hazardous material identification is described in the following section. It consists of a literature study to obtain an overall picture of the status quo of machine learning applications in hazardous building material management, summarized in Paper I, as well as an empirical study on developing a hazardous material dataset and a machine learning pipeline specific for the study objectives, synthesized in Paper II and Paper III. Accordingly, the chapter begins with a description of the systematic literature review to map out the main findings and development of the field (Section 3.1), then continues with presenting the work on data assembling and validation (Section 3.2). Finally, a proposed method for structuring a machine learning pipeline is described together with the demonstration of two prediction cases (Section 3.3).

3.1. Systematic Literature Review

A systematic literature review was done to identify, select, and evaluate earlier research work on data-driven applications for hazardous material management. The method for conducting the search process was described in Paper I. Figure 3.1 illustrates a two-fold review procedure of science mapping (Section 3.1.1) and critical review (Section 3.1.2). Science mapping quantitatively measures the domain development by computing the metadata of the literature, while critical review concerns qualitative synthesis on the content of the publications. Engaging both elements in a systematic literature review is necessary for a comprehensive research appraisal. The outcomes of the science mapping offer a deeper understanding of the core research activities, conceptual and intellectual structures of the publications. Based on the highly relevant literature, a number of relevant machine learning applications for hazardous material identification, separation, and collection were highlighted.

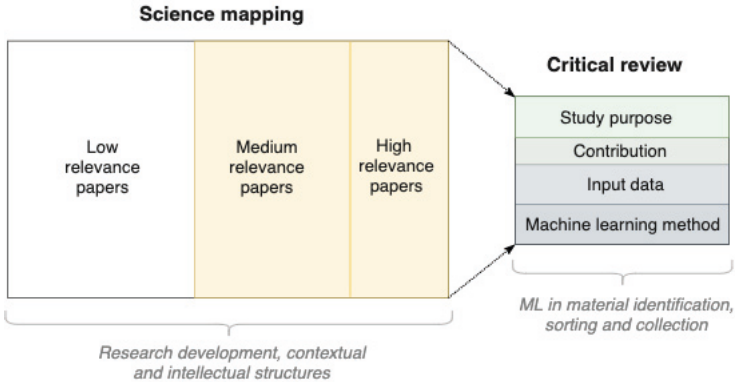


Figure 3.1 The review procedure consisted of two consecutive parts: science mapping and critical review.

To begin with, a structural review plan was designed to enable a transparent and reproducible literature search, including a pronounced search strategy, predefined assessment criteria, and information analysis tools [97]. Online generic databases were used for the initial search, then complemented with the snowball search on the secondary literature from citation lists of the acquired documents. The goal for the initial search was to screen as many relevant articles as possible, while the complementary search was undertaken to refine the search relevancy. Two search engines – Web of Science and Google Scholars – were used as they contain interdisciplinary publications with a broad timeframe and language choices. However, they curate information with distinctive logic. Web of Science is an expert-based database containing publisher-neutral, peer-reviewed academic papers, whereas Google Scholars is a robot-based document retrieval platform featuring high adaptability in full-text searches on any document type. The differences imply that the search results from Web of Science may be more rigorous and reliable, yet it may overlook heterogenous but relevant documents. Hence, using the exact search term on Google Scholars again can avoid the risk of missing up-to-date articles.

The search terms – "hazard", "artificial intelligence (AI) or machine learning", and "building" – marked with wildcard expression combined with Boolean operators were used for literature query on Web of Science. The search results returned English-based literature, including articles, proceeding papers, reviews, conference papers, with the search terms in their title, abstract, author keywords, and frequent-occurring reference keywords without a specific timeframe. The extensive search scope was preferred over the traditional keyword search to have an overarching examination of the document's relevancy. After obtaining the first batch of literature, an iterative refining process on the search terms was undertaken to exclude the irrelevant search results. Then the refined search phrases and the same search process were

repeated on Google Scholars to identify the articles left out from Web of Science. Appending these documents and the secondary literature from citation searches to the initial literature pool, the basis for the science mapping was outlaid. An overview of the search scope and process involved in the literature retrieval can be referred in Table 1 in Paper I.

After that, a quick paper screening on titles, keywords, and abstract was performed to level relevant literature based on the following assessment criteria: (1) the high relevant group contains the study scope about hazardous building materials AND AI/machine learning, (2) the medium relevant group involves either hazardous building materials OR AI/machine learning and (3) the low relevant group concerns the umbrella terms of AI in the architecture, engineering and construction industry, CDW management, and circular economy. The purpose of the article clustering was to delimit the core literature to the interdisciplinary subject for further content analysis in the second part of the literature review.

3.1.1. Science Mapping

Science mapping, a kind of bibliometric analysis describing the research evolution in the study domain, was implemented on the acquired documents [98]. The bibliometric analysis provides insights on the research centrality regarding research terms, key references, and the associated citation networks in the hazardous materials field. Through computing the metadata of documents into the R programming language-based bibliometric library Biblioshiny, the contextual and intellectual relationships between the publications can be visualized. Various metrics were considered during the information processing and presentation. The first metric relates to research development quantification. Through plotting the number of the research activities over the years and their thematic distribution, an integrated perspective regarding research evolution can be created. The results can be, for example, an accumulative number of publications in the field, literature proportion across disciplines, and so on.

The conceptual structure, concerning the dynamic growth of the concepts and topics, was the second metric. The co-word analysis and the word dynamic analysis show the centrality of the field, as well as the association between research terms. Lastly, the intellectual structure appertains to the shift of research paradigm between the different generations of researchers. It focuses on citation relationships between articles, presented in the formats of a historical direct citation network and a three-field plot. The former is a historiographic mapping with the time dimension, and the latter correlates authors, keywords, and publication outlets using a Sankey diagram. Combining the results from the research development, the contextual

structure, and the intellectual structure, various angles to approach the scientific landscape were presented.

3.1.2. Critical Review

Based on the results from science mapping, parts of the PRISMA principles (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) were applied to the literature classified high and middle relevancy for critical appraisal. The PRISMA statement is a comprehensive protocol evaluating the main sections of the article for critical review, including title, abstract, methods, results, discussion, and funding sources [99]. By following the structural information reporting system, the outcomes from the article comparison can be more objective and transparent. The information collected from the exercise was compiled in an excel file for prompt retrieval. In addition, extra efforts were allocated to the method section to identify machine learning techniques and input data applied in highly relevant papers. The contexts for specific machine learning applications and the limitations were underlined to answer Research Question 1. Afterward, a critical review on how these approaches can contribute to implementing the EU Construction and Demolition Waste Management Protocol was initiated and discussed. A summary of their study purposes and contributions is provided in Appendix A and B of Paper I.

3.2. Data Assembling and Validation

The method for assembling and validation of building-specific information was explored in Paper II, which is also regarded as a basis for Paper III. The critical review of the early literature suggests a few pioneering studies that employed field data and registered data for asbestos-containing material mapping in the field [58], [60]. The potential of using registered data and the inventories for hazardous waste from pre-demolition audit for hazardous material prediction is investigated in the section. Due to the lack of a hazardous material database, a hypothesis-driven data screening process was commenced with support from domain experts. The idea was to collect the past hazardous material detection records into a database of hazardous waste inventory and extract the information as data labels to train machine learning models, described in Section 3.3. Meanwhile, creating a comprehensive national building register database that encompasses available building registers and matches partial data with the hazardous material database for model training, validation, and prediction. The

rest of the national building register database will be held-out for later prediction when the final models are deployed.

This exploratory process for data assembling and validation, a point of departure for Paper II, is described in Figure 3.2 and contains three main parts: (1) Data collection for the database of hazardous waste inventories and the national building register database (Section 3.2.1 and Section 3.2.2), (2) Data matching between two data sources for a harmonized hazardous material dataset (Section 3.2.3), (3) Data validation including delineating complete observations from outliers, low quality, missing values to be able to set up a final dataset for machine learning pre-processing (Section 3.2.4). Through creating a hazardous material dataset, the possibility of assessing occurrence patterns of hazardous materials in the existing building stock can be investigated. The compiled hazardous material dataset and the proposed database assembling procedure contributes substantially to the method development of in situ hazardous material management.

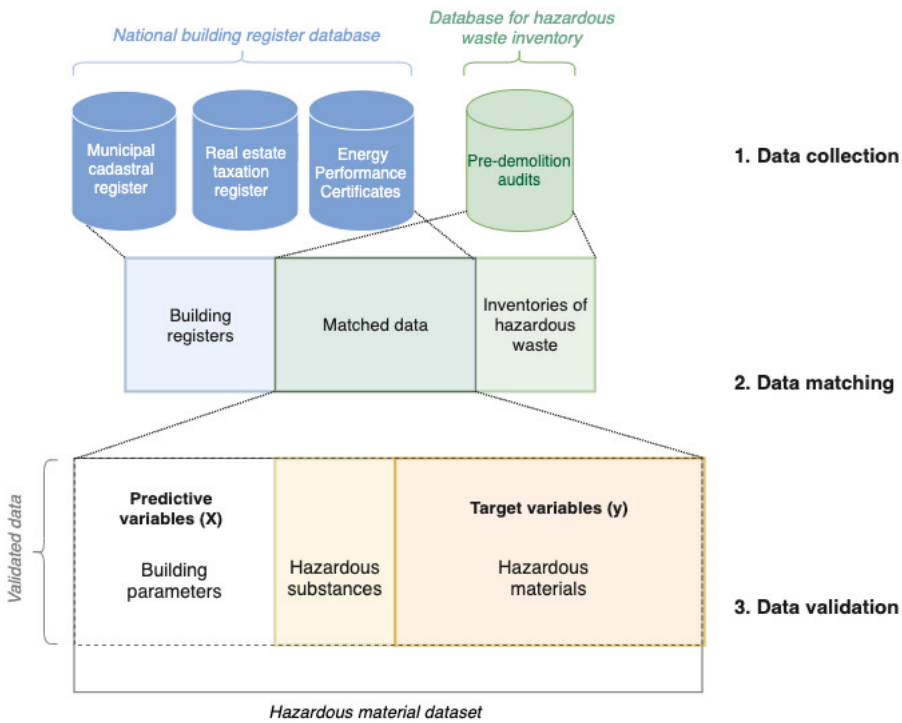


Figure 3.2 The diagram illustrates how the hazardous material dataset was created with a three-step procedure: (1) data collection, (2) data matching, and (3) data validation.

3.2.1. Database of Hazardous Waste Inventory

Inventories of hazardous waste from the pre-demolition audit of the demolished and renovated buildings are valuable data sources for hazardous material detection [32]. These field data remain unexplored in Sweden because of effort-demanding work to collect and structure the information mainly archived on hardcopy. Besides, they are scattered in each municipality as supervising the pre-demolition audit execution is the responsibility of the local authority. The first batch of hazardous waste inventories was gathered from the Gothenburg City Archive in summer 2020 for demolition and renovation permit applications applied between 2010 and 2020. Without the availability of a digital demolition database, the query was carried out manually in a general building permit register system by using the keywords "demolition", "renovation", "reconstruction", "modification", and "alteration" in the permit application descriptions. The descriptions indicate what the application concerns and which class the building is. The search returned the demolished or renovated buildings that potentially had undergone a pre-demolition audit process. After that, following an extensive document screening to retrieve pre-demolition-related documents. Then the records from the inventories of hazardous waste and the requested scanned copies were read, relevant information extracted, compiled, and reformatted into an excel sheet.

Iterating the same data collection process, the second batch of inventory of hazardous waste documents was gathered from Stockholm City Archive in summer 2021. Searching with the national real estate index in the digital, public accessible building permit database returned all the past interventions in the properties. By tracing the building history, one can obtain more certain construction years and architectural drawings for the area and floor estimation of the old buildings, which were critical variables for predicting the presence of hazardous materials [61]. This auxiliary information was critical for filling information that was missing in the inventories of hazardous waste.

The available pre-demolition audit documents from Gothenburg and Stockholm cities were stored in the database of hazardous waste inventory. The inventories can be generally grouped into four document types according to their document titles and content formats. Various data granularity and comprehensiveness of hazardous substances and materials were recognized, of which reports, protocols, control plans, and demolition plans were listed in descending order. Detailed inventories, including reports and protocols, contain records of hazardous material detection, whereas simple inventories like control plans and demolition plans only keep information at substance level, and the scope of the investigation was not defined. Reports are the most thorough documentation of an inventory,

following the detailed list of hazardous material inspections from the Swedish Circulation Council in Construction's guidelines (Kretsloppsrådets riktlinjer). Reliable detection records from lab test samples were documented by environmental consultants for complicated buildings, for instance, multifamily houses, schools, industrial buildings, and so on. Usually, the information about the investigated buildings was provided. Contrary to reports, protocols derived from the municipality template comprise a list of binary options for hazardous materials and their amount. The surveyed building parts and the scope of the investigation are indicated for all building classes.

On the other hand, control plans are designed specifically for single-family houses or simple buildings. The tabular format contains information on primary hazardous substances, i.e., asbestos, PCB, CFC, mercury. Demolition plans are required documents for demolition permit applications for general control purposes, in which free text is used to describe the detection of hazardous substances or materials, but it is hard to determine the investigated scope and actual area. Considering the distinctive data quality and information availability, reports and protocols were preferred over control plans or demolition plans if both documents were available for the same buildings.

3.2.2. National Building Register Database

Central authorities own and maintain national building registers such as real estate data, energy measurements, and population statistics. Compiling generic building registers can facilitate information comparison and supplementation across data sources. The method for merging nationwide datasets referred to Johansson et al. [31] was executed with GIS Feature Manipulation Engine from the Safe Software [100]. Three data registers were assembled into a national building database – the Swedish real estate taxation register and the municipal cadastral register [101], as well as the Swedish EPC register [102].

First of all, the Swedish real estate taxation register and the municipal cadastral register were requested from the Swedish Cadastral and Land Registration Authority. The Swedish real estate taxation register, as shown in Appendix A Table A2 in Paper II, was originally derived from the Swedish Tax Agency, where building ages and floor areas are kept updated. While the municipal cadastral register, shown in Appendix A Table A3 in Paper II, contains similar data but is reported by each municipality separately for all types of buildings. Apart from the basic building information, detailed building usage and legal status are provided.

Subsequently, the EPC data were requested from the Swedish National Board of Housing, Building, and Planning. Since the national EPC

regulation was implemented in Sweden in 2007, the property owners are obliged to obtain EPC for buildings prior to a sale issued, with rent purposes, frequently visited by the public, or newly built [103]. Yet, in some cases, a joint EPC may be issued to similar and adjacent buildings due to collective energy use measurement. The issue of EPC is valid for ten years and requires renewal after that; thus, the updated EPC was released in 2017. Over the years, more than 90% of the Swedish multifamily houses have EPC [31]. Since the renovation and demolition building permits were collected for the study spanning over 2010-2020, both early and renewed EPC were obtained. An overview of the EPC concerning building features is illustrated in Appendix A Table A1 in Paper II.

The national building database was constructed at the building level to match the observations in hazardous waste inventories; thus, the municipal cadastral register became the base layer of the database. However, key variables such as area, construction year, renovation year, and value year of this register have not been updated in the recent ten years and are reported incomplete by Johansson et al. [31]. On the other hand, the most reliable data source was the real estate taxation register. Yet, the registers were at the value unit level, which, in most cases, follows the property level. Therefore, the first step was to merge the municipal cadastral register and the Swedish real estate taxation register using the national real estate index as the key. This implied that duplicates were introduced when a building was matched with the real estate taxation register more than once. On the contrary, a building can also share value units with other buildings. Therefore, two attributes – the number of value units for a property (AntalVardeEnhet) and the number of properties for a value unit (AntalFastigheter) – were created namely to track the duplicates and the shared value units at the building level.

In addition, the total number of possible matching relationships for all building classes, presented in Table 3.1, were also included in the database as auxiliary information to control data matching quality between the hazardous waste inventory database and the national building register database, as well as add another dimension for data analysis. Based on the combination types, the matching uncertainty was investigated between building classes in the real estate taxation register for the major economic regions in Stockholm, Gothenburg, and Malmo, Sweden [104], shown in Table 3.2. It was found that small houses have the highest proportion of the 1 - 1 matching relationship, while warehouses, industrial buildings, and production buildings have a high number of other relationships, i.e., 1 - n, m - 1, and m - n. School data was lacking from the real estate taxation register due to tax exemption. Having a high percentage of other relationships indicates complex properties, leading to uncertain matching.

Table 3.1 The total number of possible matching relationships.

Combination	Description
1:1	One property belongs to a single value unit
1:n	One property belongs to several value units
m:1	Several properties belong to one value values unit
m:n	Several properties belong to several value units

Table 3.2 The total number of possible matching relationships for the real estate taxation register based on the major economic regions in Sweden.

Table	1:1 [N]	Others [N]	Others [%]
Multifamily house	24 504	3 787	13,4
Commercial building	22 098	4 108	15,7
Industrial building	1 330	327	19,7
Office	5 224	940	15,2
Warehouse	5 111	1 919	27,3
Production building	4 067	904	18,2
Single-family house	455 368	18 655	3,9

The second part of the merging related to concatenating the EPC data with the newly formed real estate taxation register and municipal cadastral register database. EPC data was structured according to the EPC index (Formular ID), which can attach to one or more properties with one or more buildings. In most cases, a building only belongs to one valid EPC index, but it may belong to one or more historical EPC indexes. By including all available EPC indexes, the analysis in Table 3.3 was conducted at the property level; namely, one or several buildings exist in the dataset. The finding shows that most EPC indexes have a 1 - 1 relationship and contain only one building. However, to prevent confusion, four attributes were added to clarify the relationship between the number of buildings and properties for each EPC index: the total number of buildings in an EPC (EPC_AntalByggnader), the count of EPC for a property (EPC_AntalEPC), the total number of properties in an EPC (EPC_AntalFastigheter), the relationships on the properties based on different combinations (EPC_Relationsklass).

Table 3.3 The distribution of the total number of possible matching relationships at the property level based on the EPC data in the major economic regions in Stockholm, Gothenburg, and Malmo, Sweden.

Combination	Description	No. properties
1:1	One property belongs to a single EPC	140 759 (80,7%)
1:n	One property belongs to several EPC	32 725 (18,8%)
m:1	Several properties belong to one EPC	826 (4,7%)
m:n	Several properties belong to several EPC	32 (0,02%)

It is critical that various granularity levels are handled with care and matched in a correct way. The current register data contain different aggregation levels for EPC index, building, and taxation value unit, of which value units are sometimes virtual units that are difficult to relate to a specific building. Nevertheless, a true match can be achieved if the EPC index is 1 - 1 and the value unit is also 1 - 1 in the case where the EPC index only has one building. However, it is often possible to puzzle and link one value unit to a building – for instance, a property with two buildings, one single-family house, and one multifamily house. The area from the multifamily house table in the real estate taxation register should be connected to the multifamily house, and the single-family house connected to the respective small house table.

3.2.3. Data Matching Between Databases

The acquired data was processed based on the data operation procedure referenced from Simon [78] and Pasichnyi et al. [29] to generate a hazardous material dataset. Table 3.4 describes six consecutive steps that were followed to ensure coherent data documentation and storage. A tabular dataset structure was created to assemble common variables between hazardous waste inventories, such as type of inventory, investigation year and scope, auditors, decontamination history, detection of hazardous substances and materials. The goal is to harmonize different detail levels of the documents without sacrificing too much of the data granularity.

Table 3.4 A data operation procedure implemented for matching and validating a hazardous material dataset.

#	Data operation	Description
0	Dataset structure	Assemble common variables between hazardous waste inventories and use “building” as an observation unit.
1	Data control	Quality control of hazardous waste inventories based on buildings construction year and investigation completeness.
2	Data conversion	Convert hazardous waste inventory records to machine-readable data formats, i.e., “nominal”, “scale variables”, and “ordinal.”, and compile them into a digital hazardous material dataset.
3	Data extraction	Retrieve the building registers from the national building database with the national real estate index from the observed buildings in the hazardous material dataset.
4	Data matching	Use the national real estate index as the key to establishing a one-to-one matching between the extracted building registers and the observed buildings in the hazardous material dataset.
5	Data revision	Check consistency between building registers and inventory data, harmonize variance and revise the matched variables.

Next, examining the eligibility of the investigated building and remove data from low-quality hazardous waste inventories. The earliest prohibition on the use of PCB and asbestos can be traced back to 1973 and 1975 in Sweden. Yet, the access to PCB and asbestos contaminating building materials from import and market circulation continued until 1982. Therefore, buildings built before the 1980s may be exposed to contaminants and are of interest for including in the hazardous material dataset. Detection records of asbestos, PCB, CFC, mercury, and radioactive concrete, due to the risk of radon, were documented as hazardous substances and hazardous building materials in a binary manner. After that, the quality and the completeness of hazardous waste inventories were controlled to prevent falling into the trap of a skew dataset. Considering the operationality of classification models, the extracted information from the hazardous waste inventories was converted into machine-readable data formats. Lastly, compiling these transformed observations into a hazardous material dataset.

To verify the interpretation results of the hazardous waste inventories, ten properties were selected for an observation validation exercise from the list of renovation and demolition projects in Gothenburg based on the diversity of inventory types, building classes, and building complexity. The same observation template and pre-demolition documents for the chosen properties were distributed within the research group. The template adopted the same structure as the hazardous material dataset that constituted general building information and detection results. After collecting individual observations, the compiled results were employed to a validation metrics, where every two individual observations were paired to a total of 6 sets of comparisons in a scoring system to understand the degree of agreement and divergence in interpreting raw data.

The validation exercise of inventory interpretation revealed the importance of using uniform documentation in the correct interpretation of pre-demolition documents and an agreeable data recording workflow. The disagreement on interpreting the missing values and uncertain investigation results, such as presumable or experience-based positive, was highlighted. According to the results of the validation matrix in Appendix A, around 72% of disagreement was shown in the results from 5 participants. However, the disagreement on variables varied: (1) the most profound agreement was on construction year, detection of asbestos and pipe insulation; (2) medium agreement was on investigation scope, detection of PCB, mercury, lighting tubes, sealants, capacitors in lamp or burner, door or windows insolation, carpet glue; (3) the least agreement was on CFC and the rest of materials. Confusion between “no detection” and “missing values” was the primary reason for discrepancies. A case-by-case check on the highly disagreeable variables was further conducted to understand the causes in detail. It was found that how the detection results were displayed played a critical role in

interpretation. If the detection results were listed explicitly in a tabular format or on a building basis, misunderstandings are less likely to happen. Yet, if an inventory concerns several buildings in complex properties such as schools, hospitals, industrial buildings, or multifunctional buildings, it tends to result in confusion or misleading interpretation. The results from the validation exercise indicate that improvement was needed to assure result comparability; thus, the entire process of documenting raw data from inventories of hazardous waste was repeated with new, more strict routines for interpretation.

The next step was to add general building information from the national building database to the hazardous material dataset. Firstly, the building registers of the observed buildings were retrieved with their national real estate index and used as the key for creating one-to-one relationship matching. Google Street View, hitta.se real estate map [105], and the GIS for regional buildings were adopted as auxiliary sources to check that the building registers correspond to the observed building. The issues for unmatched observations were classified with different match codes, such as too little information to determine the correct registers, several inventory observations shared one register, eliminated registers due to building demolition, or unmatched data. Depending on the extent of information loss, uncertain observations and observations without building registers have been removed from the dataset to be used for machine learning modeling.

As the variables such as construction year, renovation year, and area appeared in all data sources, the selection of the registers was based on value alignment to the inventory data. But if the values were not identical, the real estate taxation register and the EPC data were preferred over the municipal cadastral register considering their update frequency and information completeness. Furthermore, revised columns harmonizing the variances between data sources were generated in the database for building class, construction year, renovation year, area, and the number of floors, which will later be used as predictive variables. The purpose of creating building classes is to cluster the buildings with a similar function and typology. According to the renovation or demolition permit description, the primary usage of the building stated in the inventory data, as well as building types and building categories from national building registers, the observations were categorized into ten building classes: single-family house, multifamily house, temporary building, school, office, commercial building, production building, industrial building, warehouse, and other/infrastructure. The label of inventory types and building classes are fundamental to stratify the data subgroup for comparative analysis. The combination of building registers and inventory data constituted the hazardous material dataset that was used for data analysis and machine learning modeling.

3.2.4. Validation of the Hazardous Material Dataset

Data validation concerns data quality control and potential data stratification for predictive data analysis. These operations were carried out in Python's scientific computing libraries Numpy and Pandas [106], as well as statistical visualization libraries Matplotlib [107] and Seaborn [108]. The first part of data validation examined the hazardous material dataset's quality by evaluating data uncertainty in building registers and data completeness from inventories. The correct matching registers and less uncertain matching registers, whose real estate index, address, and parts of building parameters were consistent, were stratified. The data quality control resulted in a total of 848 observations, which became the basis for missing values and positive detection ratios computation.

Afterward, explorative data analysis was performed to understand the underlying data structure. As most hazardous substances were banned in the 1970s, the buildings built between 1900-1990 were of interest for descriptive analysis, which corresponds to the decades of massive use of asbestos and PCB-containing materials in buildings. 848 observations fulfilling the condition were visualized in terms of building parameter distribution against the detection of hazardous substances. Firstly, the normalized density plots and histograms of the building parameters were created for the observed buildings to understand their representativeness to Gothenburg and Stockholm building stocks. Then the distribution of the construction year for the multifamily houses and school buildings between the hazardous material dataset and Gothenburg and Stockholm building stocks were compared using normalized density plots to gain an overview of contaminated buildings.

In the last part of data validation, a cross-validation matrix evaluating the data quality and quantity were set up to streamline the process. By employing the metadata from the subgroup of each building class and hazardous material to Formula (1), assessment scores can be calculated to identify potential data for machine learning modeling. This evaluation has been done at individual observation and data subgroup levels. More specifically, the entire calculation process consists of three steps: (1) extracting the dummy-format detection results of hazardous substances and weighting inventory types based on level of detail. The weights from high to low in decile points correspond to reports ($r = 1.0$), protocols ($p = 0.75$), control plans ($c = 0.5$), and the demolition plans ($d = 0.25$); (2) for each hazardous material in a given building class, the number of observations in each inventory type was multiplied by the respective weight, then the values were summed up and divided by the number of observations in the subgroup; (3) a threshold (K) was introduced to evaluate the number of missing values for each subset. Sufficient data is a prerequisite for drawing

statistical conclusions. Therefore, if the number of observations is above 5% of the total observations in the dataset, denoted as 1; between 2,5% and 5% of the total observations, denoted as 0,5; less than 2,5% of the total observations, marked as 0. Implementing the cross-validation matrix to each data subgroup paired with hazardous materials and building class, the promising targets for the prediction can be ranked out.

$$y = \frac{(Ir \times nr + Ip \times np + Ic \times nc + Id \times nd)}{n} * K \quad (1)$$

y = Assessment score [0 - 100].
I = Inventory type for weighting the individual observation. I = 1,0 if is the report (r), I = 0,75 if is the protocol (p), I = 0,5 if is the control plan (c), and I = 0,25 if is the demolition plan (d).
n = The number of observations in the subgroup [$0 < n$].
N = The number of the observations in the entire dataset.
K = Number of observations enough for statistical operation. K = 1,0 if $n >=$ (0,05*N), K = 0,5 if (0,025*N) =< n < (0,05*N), K = 0 if n < (0,025*N).

3.3. Machine Learning Pipeline

Based on the validated hazardous material dataset, machine learning models were created to explore the prediction possibility of hazardous materials. The objectives of Paper III are to develop a machine learning pipeline that can process the data, feed them into training models, and generate prediction results for evaluation. The structure of a machine learning pipeline is illustrated in Figure 3.3 and described in detail in the following subsections. The entire pipeline constituting data processing (Section 3.3.1), model development (Section 3.3.2), and result interpretation (Section 3.3.3) was created to investigate Research Question 3. Beyond the focus on prediction accuracy, more feasible aspects regarding model tuning and generalization were also explored. This could be, for instance, the number of optimal features and input data for specific prediction and performance evaluation between classifiers. Finally, the last part of the pipeline emphasizes insights disclosure of the black box through generating hypotheses based on domain knowledge. The entire work was completed with Python's machine learning toolbox scikit-learn. Deploying such a machine learning pipeline can scale the hazardous material prediction in the large-scaled, yet not surveyed building stock and facilitate preventive building material management.

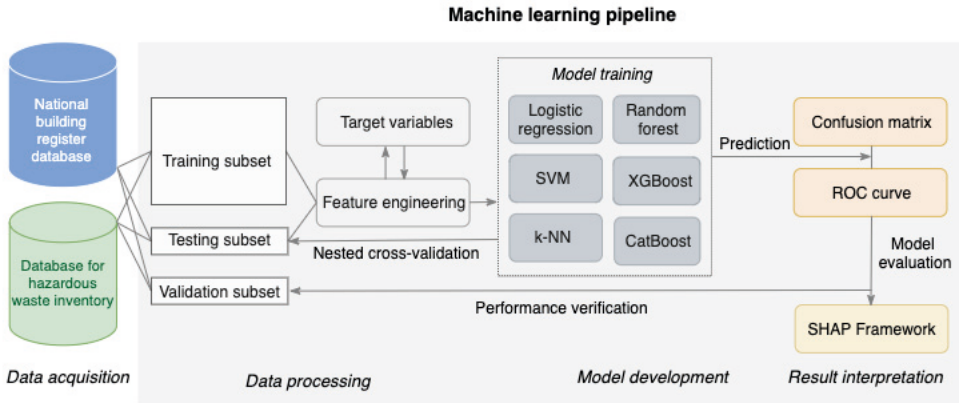


Figure 3.3 A machine learning pipeline demonstrates how data were processed and used for model development and result interpretation.

3.3.1. Data Processing

The hazardous material dataset consists of model and complementary parts, as shown in Table 1 in Paper III. The model part contains variables for machine learning modeling, such as geographics, building usage, building parameters, detection of hazardous substances and materials; while the complementary part includes auxiliary information concerning matching keys, permit description, building class, municipality building category and code, and the metadata from inventories of hazardous waste. The dataset was stratified according to building class to predict the presence of a specific hazardous material in a particular building type. By clustering the observations for similar building profiles can increase the chance of pattern identification and prevent false inference of the prediction results. The building classes with high cross-validation scores, including multifamily houses and school buildings, became the prediction targets. Then the model part was further split into 60% of training, 20% of testing, and 20% of validation subsets. This proportion for dataset partitioning was chosen for small datasets by convention. The training and the testing subsets were intended for model training and hyperparameter tuning, while the validation subset was held out for model deployment. Hyperparameter tuning allows optimizing models' performance by choosing an optimal set of parameter values for the learning algorithm [81]. Therefore, only the training and the testing subsets underwent data processing procedure, which incorporates necessary preparation before machine learning modeling, namely, data cleaning, transformation, and feature engineering.

The data transformation process started with data cleaning, where missing data and outliers were removed or replaced with mean values. This was followed by performing data transformation to optimize the performance of scale-invariant classifiers, where categorical variables were encoded to dummy variables and the numerical variables were standardized to a comparable scale [81]. Then the potential correlation between the target and predictive variables was studied using correlation plots and stepwise logistic regression. Independent variables with a high coefficient and a low p-value, which is the calculated probability that describes the likelihood of the data occurring by random chance, were selected as the initial set of potential features [109]. Then they were computed with two feature engineering techniques to identify the most relevant variables before modeling. The Recursive Feature Elimination (RFE) technique removed variables iteratively and determined the optimal number of features [110]. The results were further verified with tree-based estimators and the key features identified by RFE and Extra Trees ensemble were employed as predictive variables for model development.

3.3.2. Model Development

Several classifiers were chosen for model training considering their strengths and weaknesses for predicting on the small dataset: logistic regression, kernel support vector machines (SVM), k-nearest neighbors (k-NN), random forest, extreme gradient boosting (XGBoost), and CatBoost. An overview of the chosen non-parametric algorithms that can adapt to increasing parameters is described in Table 2 in Paper III. Introducing different algorithms can diagnose bias and variance trade-offs and adjust the model parameters accordingly [81].

The logistic regression classifiers that estimate the probability of class distribution were used as a base model to compare with the prediction performance of other algorithms. The kernel SVM classifiers, on the other hand, are high variant distant-based algorithms that linearly separate data by maximizing the data gaps on the projected hyperplane. The k-NN classifiers are instance-based algorithms that incrementally learn from data and predict the classes based on the majority vote. The random forest overcomes the drawback of overfitting tendency in decision trees and can be trained without specifying and standardizing parameters. The XGBoost with gradient-boosting allows optimization of loss function by merging weak learners with strong learners for the next prediction. The addictive models do not require model regularization and can process missing data efficiently; however, they cannot handle categorical features and need a long training time. With respect to the drawbacks, the CatBoost classifiers

were developed to process the categorical features parallelly while preventing target leakage using ordered boosting [111].

As the class imbalance between positive and negative detections was noticed during the explorative data analysis; thus, the data number of the minority class was upsampled to match the majority class for cost-sensitive learning, which implies the use of a cost matrix to calculate the total cost of misclassification by weighting the sum of the false negatives and false positives. After model training, their generalization performance was evaluated with the testing subset. The nested cross-validation, illustrated in Appendix A of Paper III, was used to obtain initial accuracy for model selection. Furthermore, the classification results were evaluated with various performance evaluation metrics, illustrated in Appendix B of Paper III. The accuracy and the recall rates in the confusion matrix were used to assess the models' performance. The accuracy entails the ratio of correctly predicted observations to the total number of observations, whereas recall (or sensitivity) implies the ratio of correctly predicted positive to all observations in actual class [81]. The Receiver Operating Characteristic (ROC) curves were schemed as a secondary performance metric, where varied discrimination thresholds evaluate the trade-off between sensitivity and specificity, namely, the true positive rate against the true negative rate [81]. Besides, the effect of data size on prediction accuracy was investigated by increasing the amount of training data. Plotting the learning curves enables measuring the bias-variance trade-off ascertaining the minimum amount of required data. In the end, the refined models were verified on the validation subset for unbiased evaluation before model deployment.

3.3.3. Result Interpretation

Despite promising prediction performance, the results from machine learning models are sometimes hard to understand by human experts. Accuracy can vary slightly in each training iteration due to the randomness nature of stochastic learning algorithms. Therefore, different visualization applications were developed to increase model transparency and enhance interpretability [75]. SHapley Additive exPlanations (SHAP) was a universal framework to explain the structure patterns determining the predicted probability [112]. Afterward, domain experts were involved in appraising whether the recognized feature importance was reasonable and coherent to the scientific hypotheses. Through the implementation of the hybrid approach, the underlying decision mechanism of algorithms can be comprehended.

4. Results

The chapter summarizes the main findings from the papers in the thesis work. It is structured into three sections corresponding to the research questions. The outcome from the systematic literature review presented in Paper I forms the background for Paper II and Paper III (Section 4.1). Paper II explains how a hazardous material dataset and the associated databases were created (Section 4.2). Validating the matched data in terms of their quality and quantify lays a foundation for the subsequent machine learning modeling. In Paper III, the prediction cases adopting the proposed machine learning pipeline were demonstrated (Section 4.3).

4.1. Emergent Applications for Hazardous Material Management

Paper I provided an overview of the existing data-driven applications for hazardous material management and highlighted the research gaps. In the first part of the literature review and science mapping, the scattered nature of the research domain was recognized in terms of their thematic distribution, and 70% of the articles were published between 2005-2020, with a peak in 2018. The major topics identified in the literature can be categorized into (1) big data, data mining, and machine learning applied in the construction sector, (2) construction and demolition waste management under the circular economy umbrella concept, (3) evaluation or remediation of the exposure risk to hazardous substances.

From a research development perspective, these topics are rarely addressed together but studied from mainly three disciplines individually – the environmental sciences and ecology, the public environmental and occupational health, and engineering. Compared to the high number of chemical-oriented references, the hazardous materials studies published in building-related journals are relatively few and accounted for only 4% of the thematic distribution among the acquired literature, presented in Figure 2 of Paper I. In particular, the distinct disconnection of hazardous material management between the use and end-of-life phases in the building lifecycle is observed. In situ hazardous material management concerns mainly concentration or emission monitoring [113]–[115], mitigation and remediation measures [96], [116],

while hazardous waste was addressed from risk management [27], [48], and source detection perspectives [117]. However, very few references concern the concept of semi-selective demolition to prevent hazardous materials from entering the waste stream and the awareness of material pollution caused by secondary contaminants. The integration between chemical and environmental engineering, and construction and building technology remains to be further developed to enable the newly-form circular economy value chain in the construction sector.

Analysis of the research front and domain evolution indicated the varied extent of research intensities between substances. According to the results of the co-word analysis, a multiple correspondence analysis measuring the similarity and variance between the Keyword Plus, the literature on asbestos and PCB-containing materials dominates the field. The identified frequent terminologies fall under two separate clusters – the diseases, risk groups, and measurement methods associated with asbestos, as well as the source materials and contamination of PCB in buildings. These results are in line with the findings from the word dynamic analysis. The accumulated keywords of "exposure" and "buildings" exceeded the safety and medicine-related terms in 2012 and gradually became the field's research focus.

Accordingly, the shift of the research paradigm is recognized in the historical direct citation network. Several bans and restrictions of asbestos worldwide came into force during the 1970s [118], the period when the research on asbestos concentration and abatement was initiated [119]–[121]. The second wave of asbestos research contains studies from 1995-2016 on the causality between the deadly diseases, i.e., lung cancers, mesothelioma, mortality, and asbestos exposure [122]–[125]. The awareness of the correlation between pollution and health was risen after that [126], [127]. To quantify the effects of asbestos-containing materials, new identification methods built upon remote sensing were developed and tested on a regional scale [56], [74], [128]. In comparison to the early research on asbestos, the studies on PCB began in 2002, and several published studies investigated PCB sources and emissions in schools, and residential buildings arose ever since [27], [115], [136], [137], [117], [129]–[135]. The relation between influential authors, keywords, and publication journals was visualized in a three-field plot (or Sankey diagram) in Figure 6 in Paper I to summarize the research magnitude.

However, a common objective is overlapping in terms of quantifying the exposure risk of hazardous materials in the built environment. Novel approaches and tools were explored to facilitate hazardous material management in correspondence to the local legal requirements. The various granularity of building classes and measurement scales, as well as a wide assortment of hazardous material in the highly relevant literature, show high empirical potential. However, their practical application to circular construction and possible synergies between different approaches remain unexplored.

Therefore, an attempt to match the content review according to the EU Construction and Demolition Waste Management Protocol [20] was made in Figure 7 in Paper I. Highly relevant literature involving all search terms was evaluated based on the first part of the protocol. More specifically, the concerned articles' research objectives, data specification, and analytic techniques were illustrated into three quantitative purposes: hazardous material identification, separation, and collection, shown in Table 4.1.

Table 4.1 An overview of data-driven applications to hazardous material identification, separation, and collection.

Applications	Techniques*	References
Identification - Remote sensing		
Identify asbestos-cement roofing	CNN, RF, Naïve Bayes, SVM, k-NN, LDFA/QDFA, Boruta	[58], [60], [138], [139]
Identification - Building investigation		
Identify the presence of asbestos materials	Cohen's kappa statistics	[45]–[47]
Assess the amount and costs of asbestos materials	Person correlation	[32]
Predict the presence of asbestos materials	Ontology / probability	[61]
Separation - Hyperspectral imaging		
Detect asbestos materials	PCA, PLS-DA, SIMCA	[62], [63]
Collection - Image processing		
Optical sorting CDW materials or minerals	CNN, SVM, MLP, PCA, RF, SVM, Decision Tree, Naïve Bayes	[64]–[68]

* Abbreviation of classifiers for different learning problems.

- Statistics: Partial Least-Square-Discriminant Analysis (PLS-DA); Soft Independent Modeling of Class Analogies (SIMCA)
- Feature engineering: linear/quadratic discriminant function analysis (LDFA/QDFA), Boruta
- Supervised learning algorithms: random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN), Naïve Bayes, Decision Tree
- Unsupervised learning algorithms: principal component analysis (PCA)
- Deep learning algorithms: convolutional neural networks (CNNs), multilayer perceptron (MLP)

Approaches for in situ hazardous material identification were found at both material stock and building levels. Mapping asbestos-containing cement roofing on a regional or national scale was achieved through combining remote sensing, field registers, and machine learning techniques [58], [60], yet it can only estimate a single, visible type of asbestos-containing material. Other approaches at the building level made use of building surveys [45], [47], demolition databases and pre-demolition inspection reports [32], and asbestos diagnosis and product descriptions [61]. However, as stated before in the former research, these studies have not adopted machine learning techniques for prediction. Nevertheless, the progress of probability estimation underpinned by statistics and ontology-based studies and results can be used to benchmark the positive detection rates in Paper II and verify the prediction results and influential

features in Paper III. Moreover, the hybrid method of integrating optical asbestos images and machine learning techniques for waste separation [62], [63] and collection were proved successful in previous research [64]–[68]. Various supervised, unsupervised, and deep learning classifiers show promising results for hazardous materials prediction and detection. Nevertheless, these pilot projects remain nascent and lag a broad uptake in the CDW management industry. This provides room for exploring the information from hazardous waste inventories and leveraging it as data to predict the potential presence of hazardous materials in the building stock.

4.2. Hazardous waste inventories as a Basis for Data-driven Analyses

To tackle the unaddressed research gaps, Paper II and the appended conference paper present a case study on extracting information on the presence of hazardous materials and components from hazardous waste inventories from renovated and demolition buildings in the city of Gothenburg. Later the data collection work extended to also include inventories from buildings in the city of Stockholm to increase the data size. The focus of the efforts was put into the data mining tasks to handle the unstructured building-specific data and underdeveloped register management. First and foremost, creating the database of hazardous waste inventories and the hazardous material dataset is a pilot task. The findings show that unstandardized inventory document types, varied experience levels of auditors, and investigation scopes challenge the development of a harmonized dataset structure. The different detail levels of inventories and their data reliability were described in Section 3.2.1. Hence, the primary data structure was adopted from the municipal hazardous waste inventory protocol [140] and then slightly modified based on the hazardous material list published by the Swedish Circulation Council in Construction (Byggsektorns Kretsloppsrådet) [141]. Additional attributes regarding the auditing year and building parts, coverage of hazardous materials, as well as building decontamination history were included in the dataset as extra parameters for quality leveling of the observed buildings.

Screening the hazardous waste inventories in Gothenburg and Stockholm cities resulted in 906 observations. After excluding ineligible buildings or buildings with a lack of information about building class or construction year, 848 observations remained, of which 85.0% of the observations came from reliable and complete data sources such as reports or protocols. Both inventory types investigated a wide range of hazardous materials with fewer missing values than was the case in control plans or demolition plans. To understand which building classes contain quality inventory data, the inventory types are

aggregated into a countplot in Figure 4.1. The results show that inventory data from reports are available for school buildings, multifamily houses, commercial buildings, and production buildings. On the contrary, control plans are primarily for single-family houses, and the building classes with inventories documented in protocols and demolition plans are relatively evenly distributed.

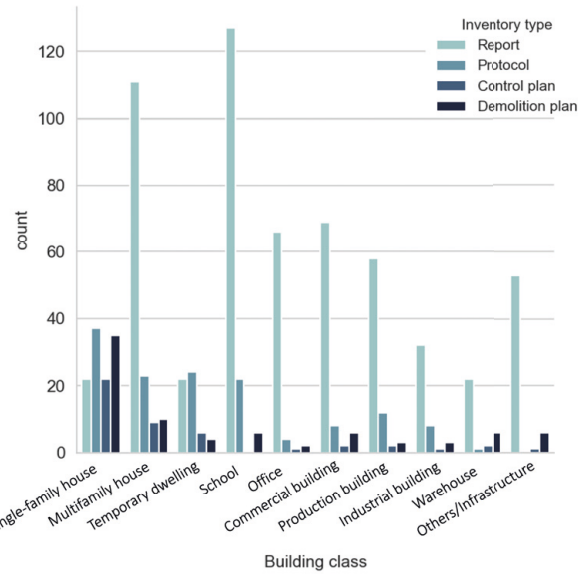
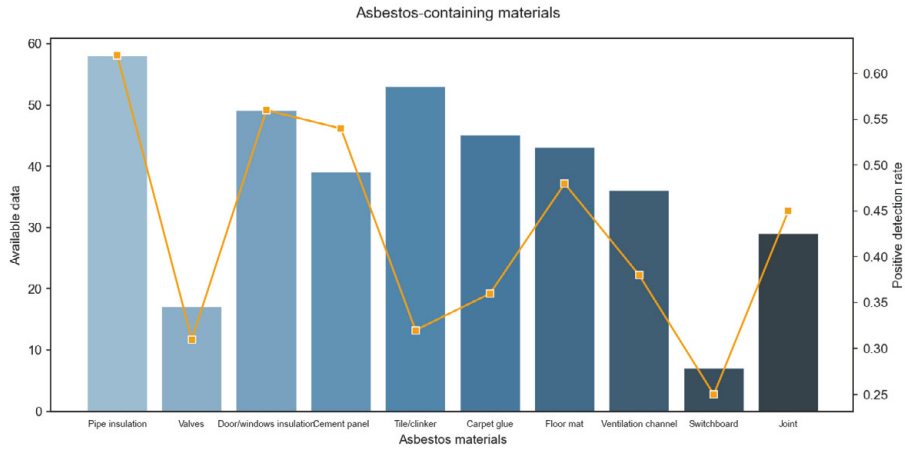
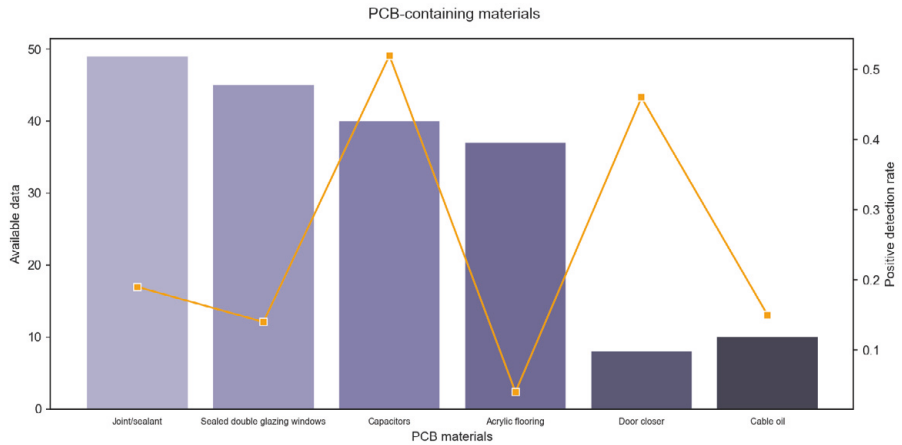
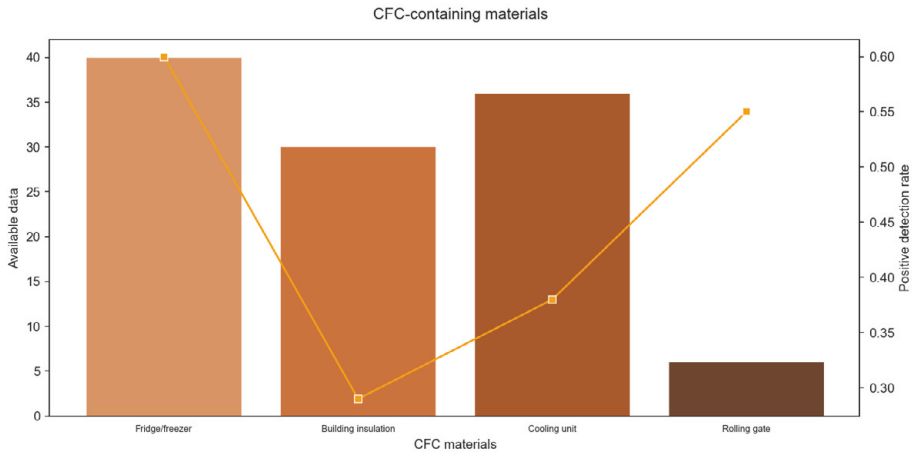


Figure 4.1 Clustering the inventory types across building classes with a counplot.

Positive detection rates and available data subgroups were visualized based on hazardous materials in Figure 4.2 for the entire dataset to highlight the potential materials for analysis and modeling. Positive detection rates are the aggregation results of the number of positive observations over the total number of valid observations, indicating contamination likelihood. The available data amounts represented as bar charts underpin the validity of the calculated positive detection rates, represented as line charts. Asbestos-containing materials with high number of detection records and positive detection rates were reported as the following: pipe insulation (r = 0,62), door or windows insulation (r = 0,56), floor mat (r = 0,48), cement panels (r = 0,54), and joints (r = 0,45). Other asbestos materials with adequate data are tile or clinkers, carpet glue, and ventilation channels. In comparison, sufficient data are obtained for PCB-containing joints or sealants, sealed double glazing windows, capacitors in lamps or burners, and acrylic flooring, of which capacitors are reported with a high positive detection rate (r = 0,52). CFC-containing fridge or freezer and mercury-containing lighting tubes are also found to have a sufficient data amount and are detected frequently.


Asbestos-containing materials
Available data
Positive detection rate
Pipe insulation
Valves
Door/windows insulation
Cement panel
Tile/clinker
Carpet glue
Floor mat
Ventilation channel
Switchboard
Joint
Asbestos materials


PCB-containing materials
Available data
Positive detection rate
Joint/sealant
Sealed double glazing windows
Capacitors
Acrylic flooring
Door closer
Cable oil
PCB materials


CFC-containing materials
Available data
Positive detection rate
Fridge/freezer
Building insulation
Cooling unit
Rolling gate
CFC materials

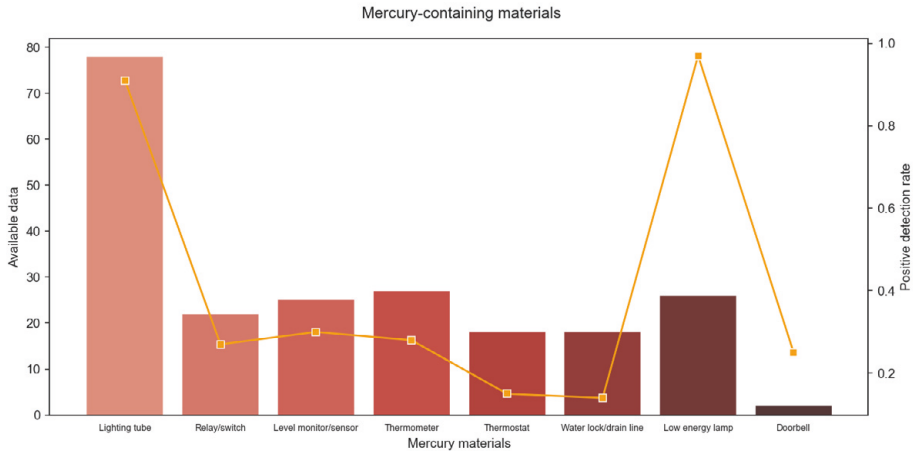


Figure 4.2 The combined bar charts and line charts show the available data amount and positive detection rates of specific hazardous materials. The illustrations from top to bottom describe the asbestos, PCB, CFC, and mercury-containing materials for the entire dataset.

Table 3 in Paper II showed that reports and protocols contain substantial detection information of hazardous materials, whereas control plans and demolition plans only have limited records on hazardous substances. Besides, the positive detection rates vary significantly between inventory types and hazardous materials when taking the sufficient data size into account ($N \geq 5\%$ of the total number of observations). Reports generally have the highest detection rates than the other inventory types, while control plans have the lowest rates. The frequently detected hazardous substances are mercury, asbestos, CFC, and PCB, listed in descending order. When it comes to materials and components, asbestos cement panels and asbestos pipe insulation have high detection rates in both reports and protocols. PCB-containing capacitors and sealants, CFC-containing fridges or freezers, and mercury-containing light tubes show similar tendencies. The variance of detection rates is assumed relating to the constitution of building classes of the observations for each inventory type. Data stratification for equivalent building types and the construction period is needed in the later work to obtain robust detection results.

Furthermore, the building parameters of the observed buildings were assessed to determine the possibility of using the detection results from the inventories of hazardous waste to represent their potential presence in the Gothenburg and the Stockholm building stocks. As the hazardous material dataset contains renovated and demolished buildings from the latest decade, the building class composition differs from the Gothenburg and the Stockholm building stock. The number of low height, small-area complementary buildings, and summer houses in the hazardous material dataset are overrepresented, whereas

multifamily houses are less representative. Also, complex properties with several buildings or buildings primarily operating a contamination-like business are enforced for environmental investigation. These kinds of buildings can, for example, be schools, industrial and production buildings. Since the aggregation level in the hazardous material dataset is individual building, they are prone to overrepresent the building stock. Considering these factors, the potential building classes from both datasets were selected for descriptive analysis.

To ensure correct inference of the results from the pre-demolition audit sampling, the value distributions of construction year were compared between the inventory data and the Gothenburg and Stockholm building registers for multifamily houses and school buildings. The results were compiled and presented as normalized density plots in Figure 4.3. The kernel density estimates show a relatively parallel distribution for school buildings. Yet, a tendency of oversampling the buildings built between 1950 and 1980 were observed both for multifamily houses and school buildings. The actual density estimates for the multifamily housing stock in Gothenburg and Stockholm are mainly spreading between 1920 to 1980, while the school building stock shows high-density distributions in the 1910s and the period of 1940-1980.

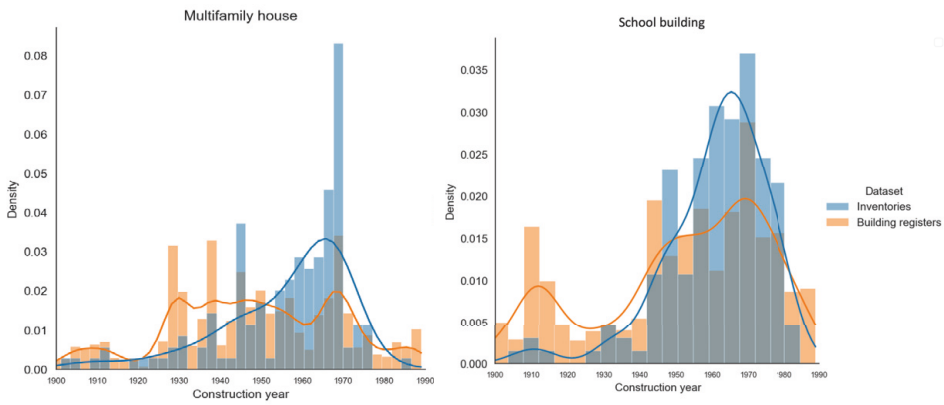


Figure 4.3 The normalized density plots are created for descriptive analysis between the hazardous material dataset and the Gothenburg and Stockholm dataset. The illustrations describe the construction year distribution in multifamily houses, including mixed-use offices and commercial spaces on the left (NI = 147, NBR = 21268) and school buildings on the right (NI = 152, NBR = 1859).

Overall, several interlinked factors challenge the use of building-specific information from hazardous waste inventories, especially data incompleteness, insufficient data amount, and heterogeneous building classes. The proposed cross-validation matrix considers these factors and pinpoints the potential

prediction of hazardous materials in particular building classes. Table 4.2 shows an overview of the calculated cross-validation scores based on Formula (1) and underlines the data subgroups whose cross-validation scores are over 90. The counted number of the high scores were described on the building class and material basis to offer an overarching picture of which data subgroups should be prioritized for machine learning modeling. Then the results were summarized into a score ranking presented in Table 3 in Paper III. According to the metadata of the assessment score, promising building classes are schools, commercial buildings, industrial buildings, multifamily houses, offices, and production buildings. The high-scoring hazardous materials aligned with the frequently investigated hazardous material list in protocols and reports. Specifically, the high potential asbestos-containing materials are pipe insulation (N= 4), door or windows insulation (N= 3), tile or clinker (N= 3), carpet glue (N= 3), floor mat (N= 3), ventilation channels (N= 3), and cement panels (N= 2), while PCB-containing materials are joints or sealants (N= 3), sealed double glazing windows (N= 3), capacitors (N= 1), acrylic flooring (N= 1). In short, both asbestos-containing pipe insulation in multifamily houses and PCB-containing joints or sealants in school buildings were found with sufficient data granularity and quantity, which made them became the prediction targets to further investigate the feasibility of using machine learning to estimate the presence of hazardous materials.

Table 4.2 The overview of the assessment scores for each building class is based on data quality and data size (N= 848, numbers in bold are the scores higher than 90).

Hazardous	**Building Class**										
Material	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**N**
Asbestos	63	**90**	80	**95**	**97**	**91**	**90**	89	41	**94**	6
Pipe insulation	74	**92**	42	**95**	50	**95**	**95**	48	0	49	4
Valves	38	46	0	**92**	0	0	0	0	0	0	1
Doorinsulation	38	**94**	41	**95**	50	**96**	48	48	0	50	3
Cement panel	36	**93**	40	**94**	50	44	44	0	0	0	2
Tile/clinker	76	**94**	40	**96**	50	**96**	46	46	0	50	3
Carpet glue	39	**93**	40	**96**	50	**93**	47	48	0	48	3
Floor mat	44	**95**	0	**98**	50	**96**	48	50	0	49	3
Ventilation channel	0	**96**	0	**98**	49	**95**	50	50	0	50	3
Switchboard	0	0	NA	50	0	0	0	0	0	0	0
Joint	0	**99**	0	**95**	50	48	48	0	0	0	2
Others	0	**96**	0	50	50	48	0	0	0	0	1
PCB	66	89	80	**93**	**96**	**93**	**90**	46	44	**95**	5
Joint/sealant	76	**91**	41	**92**	48	**94**	45	47	0	0	3
Sealed window	76	**91**	40	**94**	48	**95**	46	48	0	50	3
Capacitors	75	41	41	**94**	0	46	44	45	0	0	1
Acrylic floor	73	87	38	**94**	0	46	44	48	0	0	1
Door closer	0	0	0	46	0	0	0	0	0	0	0
Cable with oil	0	0	0	46	NA	0	0	0	0	0	0

Others	0	0	0	0	0	0	0	0	0	0	0
CFC	62	89	36	**93**	**98**	**92**	**91**	46	42	46	4
Fridge/freezer	68	**91**	39	**94**	0	44	45	46	0	0	2
Insulation	70	41	38	**93**	0	46	43	0	0	0	1
Cooling unit	36	43	38	**94**	50	48	46	47	0	0	1
Rolling gate	0	0	0	0	0	0	0	0	0	0	0
Others	0	0	0	0	0	0	NA	0	0	0	0
Mercury	62	**91**	79	**96**	**96**	**91**	**91**	89	41	**91**	6
Lighting tube	68	**91**	84	**95**	**96**	**93**	**91**	45	42	**92**	6
Relay/switch	35	40	38	**92**	0	0	44	0	0	0	1
Level monitor	38	41	38	**94**	0	0	45	0	0	0	1
Thermometer	38	44	38	**94**	0	0	44	0	0	0	1
Thermostat	37	0	38	**92**	0	0	43	0	0	0	1
Water lock	36	0	38	**94**	0	0	0	0	0	0	1
Lamp	0	48	0	**97**	50	50	0	0	0	0	1
Doorbell	0	0	NA	0	NA	0	0	NA	0	NA	0
Others	0	0	0	48	0	0	0	0	0	0	0
N	0	15	0	27	5	13	6	0	0	4	

* The representation of the building class: C1 (Single-family houses, N=116), C2 (Multifamily houses, N=153), C3 (Temporary dwellings, N=56), C4 (Schools, N=154), C5 (Offices, N=72), C6 (Commercial buildings, N=85), C7 (Production buildings, N=75), C8 (Industrial buildings, N=44), C9 (Warehouse, N=31), C10 (Others/Infrastructure, N=60).

4.3. Risk Assessment Using Machine Learning Methods

Paper III is a successive study of Paper II that features applied machine learning model development. A machine learning pipeline to perform the hazardous materials prediction tasks was developed and tested with two promising hypotheses from the cross-validation matrix. To explore the underlying data patterns, the following tasks were defined to determine (1) prediction accuracy between classifiers, (2) the minimum number of observations for reliable prediction results, (3) influential features for specific prediction models. The acquired results are intended to evaluate the possibility of adopting a new approach for risk assessment of the remaining hazardous materials in buildings. More concretely speaking, the influential features associated with the specific hazardous material's detection in particular building classes, along with the machine learning algorithms that have the optimal performance, were intended to be identified.

Previous literature indicated prevalent hazardous materials exposure in residential [32], [45], and school [115] buildings and quantified with high detection rates. The normalized stacked density distributions for asbestos-containing material detections in multifamily houses in Figure 4.4 and the PCB-containing material detections in school buildings in Figure 4.5 are presented.

In the binary classification, 0 entails negative detection, and 1 represents positive. The figures showed that the positive detection likelihood dropped in both cases when the asbestos and the PCB were banned in the 1970s. For asbestos-containing materials in multifamily houses, the density distribution patterns of pipe insulation, valves, and door and windows insulation were alike. On the other hand, the positive detection likelihoods of PCB-containing materials in school buildings were lower, and no particularly noticeable patterns were identified between materials.

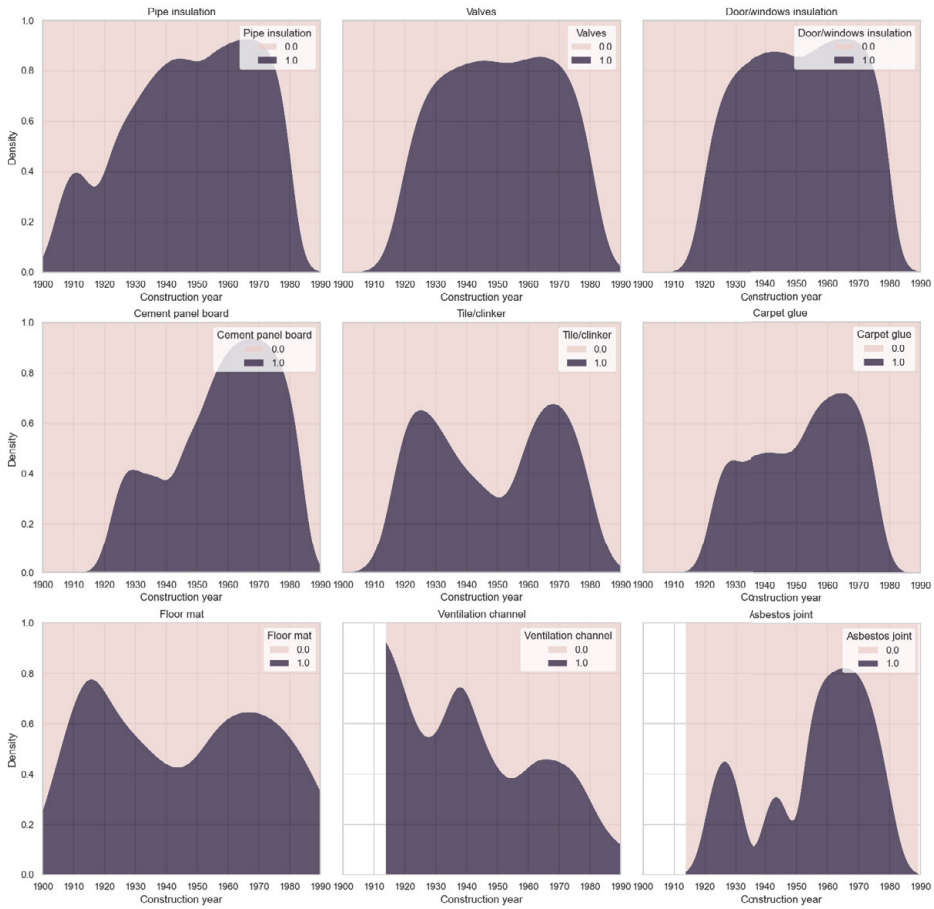


Figure 4.4 The normalized stacked density distribution shows the likelihoods of positive detection for asbestos-containing materials in multifamily houses.

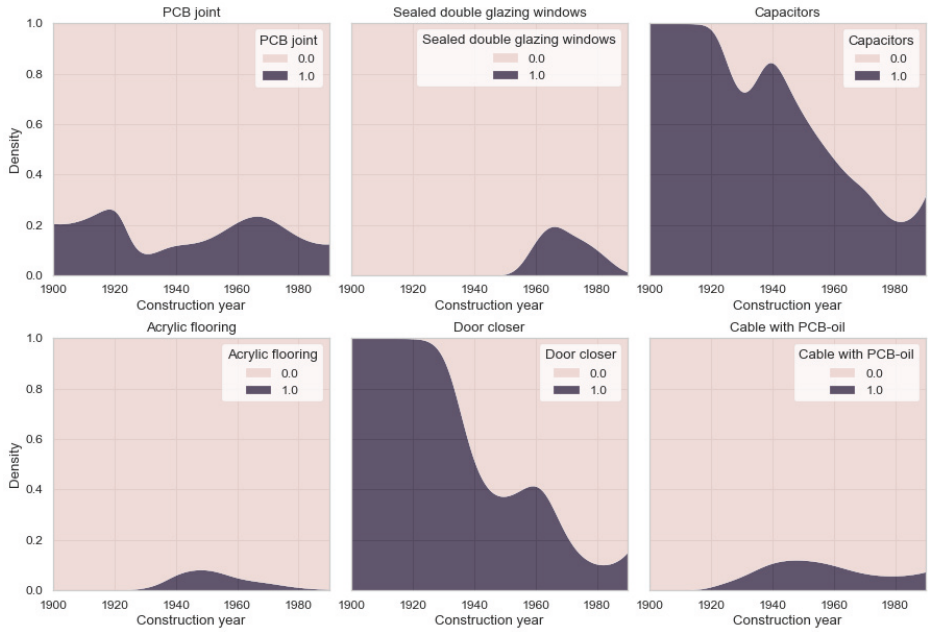


Figure 4.5 The normalized stacked density distribution shows the likelihoods of positive detection for PCB-containing materials in school buildings.

In the subsequent feature selection, the Recursive Feature Elimination (RFE) algorithm and the Extremely Randomized Trees Classifier (Extra Trees) were used to identify preferable features and the number of features. The findings from cross-validation accuracies showed that the optimal numbers of features for predicting asbestos-containing pipe insulation and PCB-containing joints or sealants were seven and four, respectively, illustrated in Figure 4 in Paper III. The chosen variables from RFE were inserted into the Extra Tree classifier for feature importance ranking, visualized in Figure 5 in Paper III. Construction year, floor area, renovation year, and the number of stairwells and floors are the key features for asbestos pipe insulation prediction in multifamily houses, whereas floor area, construction year, balanced ventilation, and renovation year are crucial for PCB joints or sealants prediction in school buildings.

By employing the input features to the machine learning pipeline, the prediction performance of the selected classifiers were evaluated. The results of the confusion matrix are summarized in Table 5 in Paper III and show that 74% of the average accuracy for asbestos pipe insulation prediction was obtained, whereas the average accuracy for PCB joints or sealants prediction reached 83%. In addition to accuracy and recall, Table 4.3 presents additional performance evaluation measures of the cross-validation F1 score and the area under the ROC curve (AUC) for model selection. F1 score considers the precision-recall tradeoff by weighting their average values. A F1 value close to

1 entails a high harmonized mean of precision and recall rates. On the other hand, AUC measures how well the model separates the classes and is usually used for evaluating class balanced classification.

Table 4.3 The cross-validation accuracies between classifiers are presented for the predictions of asbestos pipe insulation and PCB joints or sealants.

	Asbestos pipe insulation		**PCB joints or sealants**	
	F1 score*	AUC**	F1 score	AUC
Logistic regression	0,55	0,62	0,67	0,86
Kernel SVM	0,50	0,79	0,66	0,83
k-NN	0,73	0,82	0,81	0,78
Random forest	0,89	0,92	0,95	1,00
XGBoost	0,89	0,88	0,90	1,00
CatBoost	0,84	0,96	0,95	1,00
Average	0,73	0,83	0,82	0.92
Average (tree-ensembled)	0,87	0,92	0,93	1,00

* F1 score = 2 (REC * PRE / (REC + PRE))
**AUC = TPR / FPR

The tree-ensembled classifiers, including the random forest, XGBoost, and CatBoost classifiers, had notably better performance than other models in both prediction cases. In general, cost-sensitive learning that handles the class imbalanced classification performed well on the resampled dataset. The findings also showed that the random forest and the XGBoost classifiers performed the best among all classifiers in predicting asbestos pipe insulation with F1 scores 0,89, while the random forest and the XGBoost classifiers obtained optimal F1 scores 0,95 in predicting PCB joints or sealants in school buildings. To verify the prediction results, the ROC curves were plotted to illustrate the true positive rate against the false positive rate at various threshold settings. The average AUC of the asbestos pipe insulation prediction ranged from 0,83-0,92, while the average AUC of the PCB joints or sealants prediction was obtained between 0,92-1,00.

After that, learning curves were schemed to diagnose the bias and variance problems by plotting the accuracy rates along with the increasing number of training data, presented in Figure 7 in Paper III. The findings showed that satisfactory prediction results could be obtained with 100 training datapoints for asbestos pipe insulation prediction and a minimum of 50 training datapoints for PCB joints or sealants prediction. With increasing data, the gaps between training accuracy and validation accuracy decreased and gradually reached a balance bias-variance trade-off. However, the tree-ensembled classifiers had a tendency of overfitting; collecting more training data or increasing the degree of regularization can address the issue.

Lastly, contributing features and their importance to the model outputs were summarized using SHAP values, presented in Figure 8 and Appendix C in Paper III. For predicting asbestos pipe insulation in multifamily houses, the primary features identified by XGBoost and Catboost classifiers were construction year and floor area. Other influential features were, for example, renovation year, the number of stairwells, and apartments. Interpreting from the feature values, the multifamily houses built close to the 1960s-1980s with a large floor area are more likely to have asbestos-contaminated pipe insulation. On the other hand, the most crucial features for PCB joints or sealants prediction in school buildings were construction year, balanced ventilation, and floor area. Combining the feature values from both XGBoost and CatBoost models, it became apparent that school buildings built in the later era with a balanced ventilation system, large floor area, and no installed balanced ventilation system with heat exchanger are more likely to have PCB-contaminated joints or sealants.

5. Discussions

In this chapter, critical aspects of the hazardous material dataset creation and the development of a machine learning pipeline are discussed. The limitations of the data have been briefly addressed in Paper II and Paper III. Building upon that, a more comprehensive discussion about data interoperability and representativeness related to building-specific information enrichment will be presented and discussed (Section 5.1). The next part (Section 5.2) concerns the result implication of the prediction outcomes and the possibility of scaling the method to the building stock in other regions or nations. Finally, the viability of developing a digital pre-demolition protocol and its benefits are discussed.

5.1. Limitations of the Data

The robustness of the prediction results and performance of the machine learning models to predict the presence of hazardous materials in the not yet investigated building stock are highly dependent on input data. Therefore, the certainty and the characteristics of the training data are critical factors that must be assured. By delineating the data usefulness boundary, the hypothesis for the predefined prediction scope can be formulated and verified accordingly. The section discusses the data quality (Section 5.1.1) and data interoperability (Section 5.1.2) between generic building information and building-specific information, as well as the data representativeness of the compiled hazardous material dataset (Section 5.1.3).

5.1.1. Data Quality

One should be aware of the risk of mining data from hazardous materials inventories from demolished or renovated buildings. Firstly, incomplete inventories and missing values can cause misinterpretation in data transformation. It is unclear whether the suspected building materials are nonexistent or not investigated when it comes to protocols, control plans, and demolition plans. These uncertainties are derived from fragmentary building documents in the desk study, inaccessibility to the building

components, and non-systemized documentation of the investigation [32]. The “gray zone” of the auditing boundary has not been explicit, and “presumable terms” of the ambiguous investigation results lead to confusion between missing data and negative detection.

Next, the assessment scope may not represent the entire buildings as hazardous waste inventories for renovation often concerns simply the reconstructed or modified parts. Thus, using the partial detection results for model training may exclude other potential building materials and cause bias. Also, various experience levels of auditors and their inspection process can cause uncertainty of the inventory results. Sampling analysis in reports is performed by qualified consultants or environmental experts for complicated buildings; however, the investigation extent and quality of the inventories without conducting sampling may need to be further clarified. Wahlström et al. [2] pointed out that the current hazardous waste inventories focus on identifying in situ hazardous materials, and the mapping of the wastes is not fully implemented or estimated accurately. In order to reuse and recycle material-specific fractions, it will require establishing the backward information loops of hazardous waste to audit the actual identified amount and types of materials after the demolition.

Finally, various update paces between registers were noticed in data matching. Multiple construction years, renovation years, and floor areas registers from different sources hinder the exact matching between registered and inventory data. Accordingly, the GIS map of building footprint and height was introduced to enable reference value selection for floor area and number of floors variables. All available versions of EPC data and their approval dates were also used to assure correct data retrieval. Accounting for all these biases in data, a universal model that can be used to replace hazardous waste inventories will never be possible to be developed. Just as no models ever have a complete account of reality. However, the target of this research is to provide predictions that can guide pre-demolition audit investigations and assist decision-making for potential risk assessments.

5.1.2. Data Interoperability

Data interoperability relates to coupling the national building register database and the database of hazardous waste inventories. The bottom-up approach of adding building-specific information, such as socio-economic information [58], [73], spatial information [31], [57], [60], and building measurements [6], [90], to enrich the building database has been explored by former research. Several barriers were reported in integrating the distributed and heterogeneous field data with the standardized registered

data, above all, resource-demanding extraction and transformation of non-digitalized data [31], as well as time-consuming processing of non-standardized data [8]. Consequently, the data-enrichment process is manual and requires constant quality control from domain experts to assure correct building observations and documentation interpretation [6]. Similar problems were reported when collecting and exploiting demolition-related information [52], which has been experienced in this study. The incompatible interface between hazardous waste inventories and building registers challenges the integrity and consistency of the building information. It is, therefore, hard to determine which buildings have been investigated due to the lack of building keys corresponding to the building registers. Aside from that, the correctness of building information from inventories is hard to evaluate when they conflict with building registers.

Other risks include matching records from inventories and building registers for creating the hazardous material database, as well as merging multiple building register sources for establishing the national building register database. Matching data requires many relationships between datasets, yet missing information in both building registers and pre-demolition inventories entails the uncertain matching results that may cause by wrong data retrieval. To address the issue, a matching code was created to examine the matching certainty for the joined dataset in the case the observed buildings were modified or deconstructed. The matched data with a one-to-one relationship accounts for 69% of the hazardous material dataset (N=848) after removing the observations that lack building classes or construction year. Nevertheless, circa 12.7% of the observations from complex properties can hardly be distinguished, or the lack of complete information from either registers or inventory data (1.3%). Another 7.1% belong to specific matching at the property level, yet uncertain at building level because of nearly identical building parameters. On top of that, the information in the registers may be eliminated after buildings are demolished (4.2%) or renewed to reconstructed buildings (0.5%). As the extent of building information in the inventories varies, part of building parameters may be unmatched (2.7%) or have one available register information for several inventoried buildings on the same property (1.9%). It also happens in the opposite way that an aggregated inventory represents more buildings on the same property (0.6%).

Moreover, various aggregation levels were used in different registers, which makes the extensive data merging challenging. For example, value unit is used in the real estate taxation register, building unit in the municipality cadastral register, EPC index (Formular ID, a mix of property and building level) in the EPC data. To deal with the many-to-many relationship, additional attributes, such as the total number of matching relationships, the number of value units for a property, and the number of

properties for a value unit, were created to control the merging uncertainty and duplicates. In short, the data matching and merging process could be automated if the register records are connected with the identical units, maintained consistently, and made accessible for pre-demolition audit desk study and documentation if needed. It would also be advantageous if the documentation from the hazardous waste inventories were standardized, digitally stored, and available in a national database.

5.1.3. Data Representativeness

Data representativeness, which is a prerequisite to scale up the prediction results and developed models based on the hazardous material dataset, is another critical issue. The risk of data bias was highlighted from the previous studies regarding building sampling methods and details of inspection [32], [45]. For instance, Govorko et al. [45] analyzed self-assessment detection reports from the developed mobile application, and Franzblau et al. [32] evaluated the contamination risk in abandoned residential dwellings from a municipality's online demolition database and their pre-demolition inspection reports. Due to various investigation details in each type of inventory, the detection rates for specific hazardous materials in prevalent building classes in reports and protocols have been seen to be overrepresentative, while those in control plans and demolition plans were underrepresented. Consequently, the question is whether the detection rates of hazardous materials from the observed buildings can represent their potential occurrence frequency in the entire building stock. The self-sampling bias also applies to the work since the data were from renovated or demolished buildings. Possible reasons for deconstruction, such as decontamination or poor building conditions, cannot be excluded, which might not be the case for the rest of the building stock.

The size of the dataset should also be considered, given the risk of result amplification from small datasets. Therefore, assuring correct interpretation of the inventories and attaining a high degree of matching with the building registers becomes extremely important when data points are few. To judge whether the data size is sufficient for drawing analytical conclusions, one can refer to the previous works investigating asbestos in residential buildings. Franzblau et al. [32] estimated the amount, type, and decontamination cost of asbestos materials in 605 observations, while Govorko et al. [45] analyzed the occurrence frequency and conditions of asbestos materials in 702 observations. In the prediction with supervised learning classifiers, von Platten [6] used the logistic regression and the SVM algorithms for building class and characteristics prediction with 512 observations, and Cha et al. [54] employed the random forest algorithm for

demolition waste generation prediction from 784 hybrid residential and commercial buildings. The hazardous material dataset with 848 observations seems comparable to former studies, yet compiling detection records from several building classes requires further stratification and resampling. Clustering the building class for specific hazardous materials resulted in small class balanced data subsets with around 100-200 observations. Although high accuracies were obtained, tree-ensembled classifiers tended to be overfitting and required more training data to reach results with a balanced bias-variance tradeoff.

To assure representation of the prediction results, the building parameters of observed buildings were compared with the Gothenburg and Stockholm building stocks built in the same period. This criterion for dataset size was to have the distribution of training and testing subsets resemble the validation subset for a good model generalization performance. Comparing the building stock as a whole, the results from Paper II showed that a majority of the demolished and renovated projects had extremely low or high values of floor area from complementary facilities and building complexes. The renovation frequency of school buildings was higher than other buildings, making these building classes overrepresented compared to the high proportion of multifamily houses in Gothenburg and Stockholm. To prevent drawing an incorrect conclusion from a skewed dataset, the data representativeness should be controlled at the building class level to compare building properties from the corresponding data subgroups. As such, the age of the multifamily houses and school buildings from the dataset based on inventories and the building registers were compared and showed a tendency of oversampling for buildings built during a certain period. The discrepancies entail that the multifamily houses and the school buildings constructed between the 1960s and 1970s are overrepresented in the hazardous material dataset. Therefore, one should consider the possible data bais when extrapolating and interpreting the analysis or prediction results for the citywide multifamily and school building stocks.

Despite overrepresentation in the hazardous material dataset, multifamily houses and school buildings are still considered suitable building classes for machine learning prediction as they have distinct building characteristics within the same data subgroup. Not only because several asbestos and PCB materials in these building classes attained high scores in the cross-validation matrix, but also that many multifamily houses and school buildings in Sweden were built during the Million Homes program [142], meaning that they tend to be standardized in terms of building characteristics, choice of materials, indoor environment operation, and control systems [6]. The prediction results from the machine learning modeling confirmed the hypothesis and proved the possibility to identify

the underlying patterns for particular hazardous materials in these two building classes. More details are described in the subsequent section.

5.2. Discussion and Result Implications

This section discusses implications of the results related to earlier works by first assessing the analytical outcomes (Section 5.2.1) and complementing the physical interpretation at the building level (Section 5.2.2). Then the upscaling opportunities and challenges for the proposed method and its replicability in Sweden and other countries are evaluated (Section 5.2.3). Further on, the applications of the method and its benefits to stakeholders are discussed (Section 5.2.3). A bold proposal is suggested to develop a harmonized digital protocol for hazardous waste inventories to facilitate data-driven hazardous materials management.

5.2.1. Assessment of the Analytical Outcomes

The study characterizes the presence of asbestos, PCB, CFC, and mercury-containing materials in the Gothenburg and Stockholm renovated or demolished building stock. The diversity of hazardous materials in the extensive building classes has not been investigated in previous research. Not until recent years, descriptive analyses using statistical methods were performed to estimate asbestos-containing materials in the regional or citywide building stock [46], [47], [57]. At a more detailed level, the remaining asbestos quantity, substance type, abatement priority, cost, detection likelihood, and source components were specified in provincial residential buildings [32], [45].

High detection rates of asbestos were observed across studies, yet the exact values vary according to local contexts, summarized in Table 5.1. The highest value was reported by Franzblau et al. [32] that around 95% of asbestos presents in the abandoned houses in Detroit in the US, whereas Govorko et al. [45] recorded 82% of in situ asbestos in Australian dwellings. The results from our study were in line with the former research, reporting average finding asbestos in 85% of the multifamily houses and 61% of the single-family houses. The prevalent asbestos materials identified in renovated and demolished buildings in Gothenburg and Stockholm were pipe insulation (82%), door or windows insulation (81%), and cement panel (73%) in multifamily houses, and cement panel (45%), and pipe insulation (38%) in single-family houses. Similarly, floor mat (51%) and cement panels (48%) were detected frequently in Detroit in the US, but different

from common asbestos materials in switchboard (50%), eaves and soffit linings (44%) in Australian homes. In addition, 19% of the observed Swedish buildings contained PCB in joints or sealants, which is slightly more than 14% of detectable PCB-containing sealants in 70 buildings measured in Toronto, Canada [132]. Overall, the detection rates of asbestos and PCB-containing components are in agreement with those in other countries, and the small variances can be understood from different construction traditions and materials used.

Table 5.1 Summary of the detection rates of asbestos-containing materials from the actual study compared with previous research.

	Franzblau et al. [32]	**Govorko et al. [45]**	**Actual study**	
Input data	Demolition projects Detection reports	ACM app	Building registers Hazardous waste inventories	
Building class	Abandoned residential dwelling	Single-family house	Multifamily house (left) Single-family house (right)	
Aggregation level	Citywide	Regional	Citywide	
Asbestos	95%	82%	85%	61%
Pipe insulation	49%	-	82%	38%
Valves	-	-	69%	21%
Door insulation	2%	-	81%	6%
Cement panel	35%	28%	73%	45%
Tile or clinker	-	-	50%	25%
Carpet glue	-	-	53%	14%
Floor mat	10%	27%	57%	48%
Ventilation channel	55%	-	47%	7%
Joint	12%	-	67%	33%

Relevant research on hazardous material prediction was only found in studies estimating the spatial distribution of asbestos-cement roofing using supervised learning [58] and deep learning [60]. Wilk et al. [58] concluded the Pseudo-R^2, the proportion of variation in the outcome explained by the predictor variables, range between 54-76% for predicting the quantity of asbestos-cement products in the Polish building stock. In comparison to this, our study reached an average of 74-87% accuracy for predicting the presence of asbestos pipe insulation in multifamily houses, while 83-93% in average accuracy for PCB joints or sealants in school buildings. Besides, the recall rates and the proportion of actual positives identified correctly for both cases are also high, 78%-91% and 83%-92%, respectively. A summary of the accuracy rates in predicting asbestos-containing materials from the study and the previous research was described in Table 5.2. Despite disparate prediction targets and building stocks, the indirect comparison can offer insights into how well the developed models perform.

Misclassification comes with a high cost, particularly the type II error where hazardous materials are categorized as false negatives. However, cost-sensitive learning, a learning type adjusting the assumption of equal error and class distribution by most machine learning algorithms [143], was proved to successfully address the class imbalanced classification with the high F1 scores and the comparative accuracy-recall rates.

Table 5.2 Summary of the prediction accuracy of asbestos-containing materials from the actual study compared with previous research.

	Wilk et al. [58]	**Krówczyńska et al. [60]**	**Actual study**
Input data	Socio-economic data, built-up areas, field inventory of asbestos-cement roofs, historical data of asbestos plants	Aerial photographs Field survey	Building registers Hazardous waste inventories
Features	(Localization, type, quantity and amount of asbestos-cement products, building features like roof slope, type of function)	(The type of roofing, the degree of roof pitch, the type of asbestos-cement building materials)	(City, EPC category, EPC type, Construction year, Renovated, Renovation year, Floor area, Numbers of floors, Number of basements, Number of stairwells, Number of apartments, Ventilation type)
Classifiers	Random forest	Convolutional Neural Networks	Logistic regression Kernel SVM k-NN Random forest XGBoost CatBoost
Asbestos roofing	75-82%	87-89%	-
Asbestos pipe insulation	-	-	55-89%
PCB joints or sealants	-	-	67-95%

Moreover, the information concerning accuracy change associated with the training data size from the learning curves helps determine sufficient data size. Identifying the minimum required data size helps save time and effort for data collection and processing. Franzblau et al. [32] used information abstractors to extract information on asbestos-containing material from the municipality's online demolition databases and the pre-demolition inspection reports. However, this automatic information retrieval was not possible in our case due to the lack of searchable demolition databases and a mix of varied formats of pre-demolition audit documentation. This fact signifies the need to establish an online searchable data warehouse for archiving pre-demolition audits and reform the manner of documenting, collecting, and archiving hazardous waste inventories.

For predictive analysis in general, simple models with few major contributing features are preferable to prevent the risk of overfitting. This objective was achieved by plotting the optimal number of features and their importance. Feature selection was executed by considering the data availability of the prediction targets in the national building register database. For instance, schools are exempted from building taxation; thus, the available features can only be sourced from municipal cadastral and EPC data. Given various data accessibility and quality for individual building classes, merging a comprehensive building database can minimize the risk of data loss for demolished and renovated buildings in modeling.

In the previous study by Mecharnia et al. [61], the temporal descriptions of marketed products were found the most critical feature concerning the potential occurrence of asbestos materials in buildings. Using construction year as a predictive variable, the probability of the asbestos-containing adhesives, glues, coating, and sealants can be estimated for the French building stock built between 1946-1994. Other features, such as distance to the asbestos manufacturing factories and socio-economic factors, were also suggested relevant to the use of asbestos-cement products in Poland from former research [57], [58], [60]. As the machine learning models were developed for the two most populated Swedish municipalities with a high proportion of national building stock, geographical and socio-economic factors were not considered in the first iteration of the study. Our results are in agreement with the findings from Mecharnia et al. [61] and indicate that construction year is a prominent feature for predicting the presence of asbestos pipe insulation. Besides, floor area, renovation year, and the numbers of floors and apartments are secondary features to identify multifamily houses with presumed asbestos contamination. Similarly, construction year and the balanced ventilation system are the most contributing features to predict the PCB joints or sealants in school buildings. The floor area, the city where buildings are situated, and balanced ventilation with a heat exchanger can also impact the prediction results. To

summarize, these two attempted predictions show the usability of the proposed machine learning pipeline regarding screening the multiple assortments of hazardous materials and detail the predictive variables associated with their potential presence efficiently.

5.2.2. Physical Interpretation of Analytical Outcomes

The preliminary prediction outcomes verified that machine learning algorithms could capture the underlying occurrence patterns of hazardous materials in the Swedish multifamily houses and school buildings. Various evaluation metrics were used to assess the model performance, including the confusion matrix and the ROC AUC. The generated accuracy, recall, and F1 scores represented scale variables on how well the models perform in the binary classification on the testing data subsets and whether the recognized parameters contributing to the prediction results are reliable based on the SHAP values [112]. In these senses, the achieved pattern identification at the stage should be regarded as an indirect implication with respect to feature importance and tendency indication of the feature values to the attempted prediction tasks, rather than a direct diagnosis of the presence of hazardous materials for buildings as a whole. To obtain explicit classification results, the exploratory models need to be refined, trained with more data, and verified with new sets of multifamily houses and school buildings as case studies before deployment [50].

In the next steps of prediction on Gothenburg and Stockholm building stocks, the transformation of the building register dataset is required to stratify building classes based on the available municipality building category and type codes, as well as the EPC building categories and types. In other words, machine learning models are versatile to the input data; thus, the consistency of building parameter distributions between sample buildings and the building stock should be comparable. Besides, despite the previous research verifying the use of industrial construction techniques and materials in the Swedish multifamily houses built during the post-war period [4], the details of building components are lacking in the national building register database. Because of this reason, it is not possible to use the building material information, such as roofing, ground structure, building framework, and walls in demolition plans, as input features to train machine learning models. The missing information regarding this can be improved if the building structure and materials are included in the future EPC surveys or the development of a digital hazardous waste inventory protocol.

In practice, the prediction accuracy cannot reach 100% as machine learning models simplify the reality and, in most cases, use majority votes

in classification [6], [81]. To evaluate the usefulness of the identified patterns, the involvement of practitioners for needs assessment is planned in future research. The two-fold stakeholders' need analyses entail determining the requirements from domain experts and potential interests from stakeholders such as property owners, contractors, and waste handling companies. Through interviews and workshops with professional auditors, we can get a deeper understanding of exactly what type of hazardous materials to expect in different building classes and which to focus on in the prediction. This is essential to position the research values to the practical environmental investigations and examine whether the identified patterns from machine learning modeling are reasonable. After that, following discussions with the property owners, contractors, and waste handling and disposal companies, a hazard exposure assessment tool and a conceptual cost estimation framework can be designed accordingly. In this way, the risk of unexpected disruption of the renovation or demolition projects due to emergent decontamination can be mitigated.

5.2.3. Method Replicability in Sweden and Other Countries

In a broad sense, method replicability entails reproducing the proposed approach in international contexts and generalizing the deployed models on a national scale. The former concerns the readiness of the data infrastructure in other countries, and the latter deals with data representation for upscaling the models in Sweden. The CDW management maturity in the EU countries to adopt the proposed method was evaluated and briefly discussed in Paper II.

A comprehensive data infrastructure along the CDW value chain is the prerequisite to improving data availability and accessibility. In many EU countries, a legal obligation is enforced to report the amount of used and removed asbestos products in all renovation and demolition projects [34]. The information about asbestos wastes is more comprehensive than other hazardous materials, leading to a considerable number of relevant studies. For instance, Wilk et al. [57] and Mecharnia et al. [61] exploited national asbestos material databases in Poland and France. These databases become valuable sources that form the basis for applying the proposed data-driven hazardous material identification approach. With the gradual uptake of pre-demolition audits, more comprehensive risk management of hazardous material stocks, i.e., PCB, radon gas, can be expected in the near future.

Back to the Swedish context, the developed models will be deployed for estimating the probability of hazardous materials detection in the large economic regions of Gothenburg and Stockholm [104]. The hazardous waste inventories carried out between 2010-2020 have been collected from

the two largest Swedish cities, Gothenburg and Stockholm, which in some way is representative of the Swedish building stock. Two aspects are taken into account for assessing the models' upscaling potential. Firstly, regional differences in building stock composition and the sampling bias needed to be overcome by appending more observations from different municipalities. Next, the disproportion between the demolition building classes and their corresponding constitution in the building stock was recognized. Yet, the focus of the prediction was on the primary building classes in the Swedish building stock. According to the latest report of dwelling stock (2021) from Statistics Sweden [144], multifamily houses account for 51% of the five million dwellings, followed by 41% of one or two dwelling buildings, 5% of special housing, and 2% of other buildings. In terms of building registers, around 91% of multifamily houses are included in the Swedish EPC, leading to broad data coverage of this building class [31]. School buildings over 1000 m^2 are categorized as special buildings in the Swedish EPC data [145], and the exposure to asbestos and PCB are constantly monitored and decontaminated by the responsible authorities [146]. Based on this, there are sufficient reasons to prioritize these two building classes for contaminant risk screening.

5.2.4. Applications of the Prediction Method

Several opportunities and challenges exist for further developing the applications of the prediction methods. First and foremost, the principal advantage of using machine learning methods is that well-defined BIM models are not required to attain adequate accuracy in predicting end-of-life arising substances [50]. This offers the opportunity for exploiting the identified patterns to screen the suspected buildings as complementation for the existing environmental investigations. Nowadays, detailed hazardous waste inventories are only made for complicated, large-scaled, or high-risk buildings. A broader coverage of buildings is needed to prevent contaminated materials from entering the waste stream.

Accordingly, the prediction models can be incorporated into digital applications to facilitate the self-assessment of hazardous material presence and exposure risk. One of the existing solutions is the Hazardous Waste App launched by The Swedish Construction Federation in 2013, which helps practitioners identify hazardous waste and guide waste management [2]. Govorko et al. [47] also developed the ACM Check, a mobile application that targets Australian house owners to identify asbestos materials. It prioritizes material condition and disturbance likelihood and compiles the inspection results into a summary report. In the future, the initial summary reports can be potentially used for permit application and

quality control by responsible authorities, alleviating the knowledge gaps of auditing and enhancing the screening scope for private buildings. Meanwhile, the digital solution can simplify data collection for hazardous material research and facility management for the existing building stock.

In fact, the substantial challenge for deploying the prediction method lies in the accessibility of the validated data, as discussed before. The key points are therefore enabling retrieving central building registers, conducting a pre-demolition audit inventory and matching the data. The significance of establishing a routine to streamline the data collection, transformation, and compilation process in a queryable updating database was stressed in former studies. Govorko et al. [46] assessed the detection accuracy of using semi-automatic identifiers for asbestos materials compared to onsite inspection. Substantial agreement was found between the two methods and verified the implementation potential of the mobile application. Moreover, Franzblau et al. [32] explored the automatic data transformation method with data abstraction algorithms in demolition documents. By retrieving information systematically, the risk of introducing bias in estimating certain hazardous materials can be reduced.

A standardized data format and process for data collection are crucial for the smooth integration of diverse data sources [31]. Therefore, there has been a call to develop a harmonized protocol at the EU level to improve CDW data management and inventory quality [41]. The goal of developing a harmonized digital protocol is to link inventories for hazardous waste in practical demolition activities. The integrated protocol can overcome the lack of back-feeding quality control information loop and integrate regional auditing practice. The Flemish traceability management system Tracimat is the pilot project that assures hazardous waste inventory quality and attests the conformity with the waste management plan for selective demolition [41]. These documents are then appended to the tender specifications for selective demolition to guarantee that the contractors take the cost of decontamination into account. By consociating actors in the CDW value chain, the ambition for realizing a clean cycle from secondary materials will become closer to reality [2].

6. Conclusions

The thesis demonstrates how to use information from hazardous waste inventories as input data to predict potentially contaminated buildings. Two common types of hazardous materials in multifamily houses and school buildings were showcased to explore the method's potential for accurate risk assessment. This data-driven approach can be replicated in other countries to enhance in situ hazardous material screening, enabling well-planned contaminant removal and fostering recycled material purity. The contributions of the study can be understood from three perspectives – data assembling, method development, and practical implementation, described in detail below:

The study contributes to assembling a hazardous material dataset constituted of environmental information and building register data. To our understanding, this is one of few pioneer studies that compiles comprehensive hazardous material detection records from hazardous waste inventories. By validating the field data's quality and quantity, the research granularity of in situ hazardous material can be refined to specific building classes and extended to other hazardous components besides asbestos-containing materials, which are the most investigated in the former research.

The second contribution relates to method development in predicting the remaining hazardous materials in the building stock. Previous works had tried to infer the detection patterns of asbestos-containing materials using statistical approaches, which help characterize the local building stock but have limited global applications. Using machine learning techniques, the ability to swiftly adapt to new contexts can be assumed. The developed machine learning pipelines and models have promising generalization performance that can predict the possible presence of hazardous materials in not yet investigated buildings.

Lastly, the study has marked contributions to practical implementation by reviewing the procedure and format of hazardous waste inventories. The results provide suggestions for future improvement, including establishing a digital harmonized pre-demolition protocol and the workflow to streamline the data coupling process between field data and building registers. The current expert knowledge on hazardous material identification depends on the historical timeline of asbestos and PCB-containing materials and the practical experience. The study verified the general assumptions by quantifying the detection likelihood of various hazardous materials. The prediction outcomes can assist the pre-demolition audits by screening the likely exposed buildings, lowering the risk of unexpected

disruption in demolition and renovation projects due to required decontamination. Acquiring the information beforehand can support decision-making concerning semi-selective demolition and clean material recovery for end-of-life buildings.

6.1. Answers to the Research Questions

The first research question (RQ1) concerns the research front of hazardous material analytics.

RQ1 What are state-of-the-art data-driven applications for hazardous material management?

The results of Paper I show that the emergent data-driven techniques mainly relate to in situ asbestos-containing materials' identification, separation, and collection. These three objectives are fundamental for advancing the EU Construction and Demolition Waste Management Protocol implementation. Applying supervised and deep learning algorithms on remote sensing and registered data can estimate the spatial clusters of asbestos cement roofing on a national scale. On the other hand, statistical approaches were employed to characterize the detection patterns of asbestos-containing materials and predict their occurrence likelihood in the existing building stock. Performing unsupervised learning on hyperspectral images enables undestructive onsite detection of asbestos materials. Lastly, supervised, unsupervised, and deep learning classifiers were proved to succeed in the optical sorting of CDW materials or minerals.

The last two research questions (RQ2 and RQ3) involve data mining and machine learning model creation and evaluation.

RQ2 What is the potential to use data from hazardous waste inventories to assess the risk of hazardous materials in the building stock?

The results of Paper II show that information from hazardous waste inventories contains valuable detection records on multiple hazardous materials in demolished and renovated buildings. Exploiting this demolition-related building-specific information can enrich the national building database with regard to the presence of hazardous materials and substances. Accordingly, a hazardous material dataset based on the hazardous waste inventories from the pre-demolition audits conducted between 2010 and 2020 in Gothenburg and Stockholm in Sweden was created and validated. The empirical study quantified the contamination likelihood of asbestos, PCB, CFC, and mercury substances in materials and

products based on four types of inventory for different building classes. Hazardous materials of low data availability and quality in certain building classes were underlined and excluded from the dataset to be used for machine learning modeling. By summarizing the results with a proposed cross-validation matrix, the promising homogeneous data subgroups were stratified and used to delineate the prediction scope. The potential subgroups for machine learning modeling are identified as asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings.

RQ3 How accurate can asbestos and PCB-containing materials in specific building classes be predicted using machine learning models?

The results of Paper III confirmed the pattern identification of asbestos and PCB-containing materials in multifamily houses and school buildings using supervised learning algorithms. Applying the developed machine learning pipeline, the presence of asbestos pipe insulation can be predicted with 74% average accuracy and 78% average recall, while the corresponding results for PCB joints or sealants are 83% and 83%. The performances between classifiers were evaluated and verified for ordinary learning and cost-sensitive learning for class imbalanced dataset. Besides, sufficient data for good prediction performance was ascertained, preventing the additional cost of data collection and processing. Construction year and floor area were found to be substantially associated with the detection of asbestos pipe insulation and PCB joints or sealants. Additionally, the type of ventilation system also indicates the presence of PCB joints or sealants. Other features with medium impact to the prediction results include renovation year, the number of floors and apartments. In short, the fine-tuned optimal models can be used as a means of decision support when screening the potential presence of specific hazardous materials in the particular building classes.

7. Future Research

The work up until the PhD defense will expand the prediction scope to evaluate the available hazardous material building components and models' robustness in predicting the presence of hazardous materials in the Swedish building stock. To scale up the proposed approach to practical applications, three aspects will be included in future research:

- Regarding modeling, the developed machine learning classifiers require refinement to have an optimal prediction performance to the heterogeneous national building stock. This entails performing extensive hyperparameter tuning and sensitive analysis. Beyond the two showcases, various individual models will be developed for predicting multiple asbestos and PCB materials in particular building classes. More comprehensive model parameters should be tested to ascertain the potential synergies between the presence of hazardous materials. Besides the applications of deep learning algorithms, an emergent technique used frequently for CDW quantification, should be explored. Before model deployment for prediction on the national building stock, more data must be collected from another city for accuracy verification and model generalization improvement.
- Topicwise, the scope of hazardous material prediction will be extended to understand the correlation between ground sourced radon and radon emitted from radioactive concrete, where its detection patterns in the Swedish building stock have not been investigated. Through constructing machine learning prediction regressors, the buildings potentially built with radioactive concrete can be characterized. The outcome can contribute to policy measures for responsible authorities and abatement planning for building owners.
- Due to the Covid-19 pandemic, the planned stakeholder meetings for the needs assessment were postponed. The early findings presented in this thesis will be used to inform stakeholders to explain the possibilities and challenges in using the proposed method. Workshops with domain experts and property owners will be arranged to understand their needs and requests to facilitate practical work. Relevant stakeholders are property owners, authorities Swedish National Board of Housing, Building and Planning (Boverket),

auditors, contractors, and demolition and waste handling companies. Integrating their environmental-economic perspectives can improve model usability in practice. The research outcomes are estimated to be the basis for developing digital tools, such as a suggestive layout for the digital protocol and a web interface or mobile applications for hazardous waste inventories for auditors and property owners.

References

[1] Energimyndigheten and Boverket, “Underlag till den andra nationella strategin för energieffektiviserande renovering Ett samarbete mellan Boverket och Energimyndigheten,” 2016. Accessed: Dec. 21, 2021. [Online]. Available: www.boverket.se/publikationer.

[2] M. Wahlström *et al.*, “Improving quality of construction & demolition waste-Requirements for pre-demolition audit,” Copenhagen, Denmark, 2019. doi: 10.6027/TN2019-508.

[3] R. Kling, *Lönsam energieffektivisering : saga eller verklighet? : för hus byggda 1950-75.* VVS-företagen, 2012.

[4] B. Berggren and M. Wall, “Review of constructions and materials used in Swedish residential buildings during the post-war peak of production,” *Buildings*, vol. 9, no. 4, 2019, doi: 10.3390/buildings9040099.

[5] O. Nylander, *Svensk bostad 1850-2000*. Lund, Sweden: Studentlitteratur, 2013.

[6] J. Von Platten, C. Sandels, K. Jörgensson, V. Karlsson, M. Mangold, and K. Mjörnell, “Using machine learning to enrich building databases-methods for tailored energy retrofits,” *Energies*, vol. 13, no. 10, 2020, doi: 10.3390/en13102574.

[7] C. Björk, P. Kallstenius, and L. Reppen, *Så byggdes husen 1880-2000 : arkitektur, konstruktion och material i våra flerbostadshus under 120 år*. Svenskbyggtjänst, 2013.

[8] H. Yan, N. Yang, Y. Peng, and Y. Ren, “Data mining in the construction industry: Present status, opportunities, and future trends,” *Autom. Constr.*, vol. 119, no. August 2019, p. 103331, 2020, doi: 10.1016/j.autcon.2020.103331.

[9] M. Migliore, C. Talamo, and G. Paganin, *Construction and Demolition Waste BT - Strategies for Circular Economy and Cross-sectoral Exchanges for Sustainable Building Products: Preventing and Recycling*

Waste. 2020.

[10] M. Wahlström, J. Bergmans, T. Teittinen, J. Bachér, A. Smeets, and A. Paduart, "Construction and Demolition Waste : challenges and opportunities in a circular economy," Mol, Belgium, 2020. [Online]. Available: https://www.eea.europa.eu/publications/construction-and-demolition-waste-challenges/at_download/file.

[11] L. A. López Ruiz, X. Roca Ramón, and S. Gassó Domingo, "The circular economy in the construction and demolition waste sector – A review and an integrative model approach," *J. Clean. Prod.*, vol. 248, 2020, doi: 10.1016/j.jclepro.2019.119238.

[12] J. Bergmans, P. Dierckx, and K. Broos, "Semi-selective demolition : current demolition practices in Flanders," in *HISER conference*, 2017, no. June, doi: 10.5281/zenodo.817324.

[13] M. R. Munaro, S. F. Tavares, and L. Bragança, "Towards circular and more sustainable buildings: A systematic literature review on the circular economy in the built environment," *J. Clean. Prod.*, vol. 260, 2020, doi: 10.1016/j.jclepro.2020.121134.

[14] J. Hart, K. Adams, J. Giesekam, D. D. Tingley, and F. Pomponi, "Barriers and drivers in a circular economy: The case of the built environment," *Procedia CIRP*, vol. 80, pp. 619–624, 2019, doi: 10.1016/j.procir.2018.12.015.

[15] P. Villoria Sáez and M. Osmani, "A diagnosis of construction and demolition waste generation and recovery practice in the European Union," *J. Clean. Prod.*, vol. 241, 2019, doi: 10.1016/j.jclepro.2019.118400.

[16] R. Jin, H. Yuan, and Q. Chen, "Science mapping approach to assisting the review of construction and demolition waste management research published between 2009 and 2018," *Resour. Conserv. Recycl.*, vol. 140, pp. 175–188, Jan. 2019, doi: 10.1016/J.RESCONREC.2018.09.029.

[17] C. Consortium, "Construction and demolition waste (CDW)." https://www.collectors2020.eu/the-project/scope/construction-demolition-waste-cdw/ (accessed Nov. 03, 2021).

[18] European Commission, "Guidelines for the waste audits before demolition and renovation works of buildings. UE Construction and Demolition Waste Management," *Ref. Ares(2018)4724185 - 14/09/2018*, no. 4724185, p. 37, 2018.

[19] European Commision, "Waste Framework Directive," *2008*. https://ec.europa.eu/environment/topics/waste-and-recycling/waste-framework-directive_en (accessed Jul. 05, 2021).

[20] ECORYS, “EU Construction & Demolition Waste Management Protocol,” Brussels, Belgium, 2016.

[21] E. Commision, “COMMITTEE AND THE COMMITTEE OF THE REGIONS ON RESOURCE EFFICIENCY OPPORTUNITIES IN THE BUILDING SECTOR,” Brussels, Belgium, 2014. Accessed: Mar. 16, 2021. [Online]. Available: http://www.worldgbc.org/files/8613/6295/6420/World_Green_Building_Trends_SmartMarket_Report_2013.pdf.

[22] N. Kohler and U. Hassler, “Building Research & Information The building stock as a research object The building stock as a research object,” 2010, doi: 10.1080/09613210110102238.

[23] A. Koutamanis, B. van Reijn, and E. van Bueren, “Urban mining and buildings: A review of possibilities and limitations,” *Resour. Conserv. Recycl.*, vol. 138, no. June, pp. 32–39, 2018, doi: 10.1016/j.resconrec.2018.06.024.

[24] N. Kohler, “From the design of green buildings to resilience management of building stocks,” *Build. Res. Inf.*, vol. 46, no. 5, pp. 578–593, 2018, doi: 10.1080/09613218.2017.1356122.

[25] M. Rašković, A. M. Ragossnig, K. Kondracki, and M. Ragossnig-Angst, “Clean construction and demolition waste material cycles through optimised pre-demolition waste audit documentation: A review on building material assessment tools,” *Waste Manag. Res.*, vol. 38, no. 9, pp. 923–941, 2020, doi: 10.1177/0734242X20936763.

[26] S. Donovan and J. Pickin, “An Australian stocks and flows model for asbestos,” *Waste Manag. Res.*, vol. 34, no. 10, pp. 1081–1088, Oct. 2016, doi: 10.1177/0734242X16659353.

[27] M. L. Diamond, L. Melymuk, S. A. Csiszar, and M. Robson, “Estimation of PCB stocks, emissions, and urban fate: Will our policies reduce concentrations and exposure?,” *Environmental Science and Technology*, vol. 44, no. 8. American Chemical Society, pp. 2777–2783, Apr. 15, 2010, doi: 10.1021/es9012036.

[28] M. Mangold, M. Österbring, and H. Wallbaum, “Handling data uncertainties when using Swedish energy performance certificate data to describe energy usage in the building stock,” *Energy Build.*, vol. 102, pp. 328–336, Jun. 2015, doi: 10.1016/j.enbuild.2015.05.045.

[29] O. Pasichnyi, J. Wallin, F. Levihn, H. Shahrokni, and O. Kordas, “Energy performance certificates — New opportunities for data-enabled urban energy policy instruments?,” *Energy Policy*, vol. 127, no. October 2018, pp. 486–499, 2019, doi: 10.1016/j.enpol.2018.11.051.

[30] Z. Yang, F. Xue, and W. Lu, “Handling missing data for construction waste management: machine learning based on aggregated waste generation behaviors,” *Resour. Conserv. Recycl.*, vol. 175, no. April, p. 105809, 2021, doi: 10.1016/j.resconrec.2021.105809.

[31] T. Johansson, T. Olofsson, and M. Mangold, “Development of an energy atlas for renovation of the multifamily building stock in Sweden,” *Appl. Energy*, vol. 203, pp. 723–736, Oct. 2017, doi: 10.1016/j.apenergy.2017.06.027.

[32] A. Franzblau, A. H. Demond, S. K. Sayler, H. D’Arcy, and R. L. Neitzel, “Asbestos-containing materials in abandoned residential dwellings in Detroit,” *Sci. Total Environ.*, vol. 714, p. 136580, 2020, doi: 10.1016/j.scitotenv.2020.136580.

[33] M. Lewis, “Incompatible trends - Hazardous Chemical Usage in Building Products Poses Challenges for Functional Circular Construction,” in *IOP Conference Series: Earth and Environmental Science*, Feb. 2019, vol. 225, no. 1, p. 012046, doi: 10.1088/1755-1315/225/1/012046.

[34] Deloitte, “Study on Resource Efficient Use of Mixed Wastes, Improving management of construction and demolition waste - Final report,” Nantes, France, 2017. Accessed: Jan. 17, 2021. [Online]. Available: https://ec.europa.eu/environment/waste/studies/pdf/CDW_Final_Report.pdf.

[35] A. Akbarieh, L. B. Jayasinghe, D. Waldmann, and F. N. Teferle, “BIM-based end-of-lifecycle decision making and digital deconstruction: Literature review,” *Sustain.*, vol. 12, no. 7, 2020, doi: 10.3390/su12072670.

[36] C. M. Eastman, P. M. Teicholz, R. Sacks, and G. Lee, *BIM handbook : a guide to building information modeling for owners, managers, designers, engineers and contractors*. .

[37] M. Honic, I. Kovacic, and H. Rechberger, “Concept for a BIM-based Material Passport for buildings,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 225, no. 1, 2019, doi: 10.1088/1755-1315/225/1/012073.

[38] BAMB, “Buildings As Material Banks (BAMB2020).” https://www.bamb2020.eu/ (accessed Nov. 16, 2021).

[39] S. O. Ajayi *et al.*, “Waste effectiveness of the construction industry: Understanding the impediments and requisites for improvements,” *Resour. Conserv. Recycl.*, vol. 102, pp. 101–112, 2015, doi: 10.1016/j.resconrec.2015.06.001.

[40] C. Bodar *et al.*, “Risk management of hazardous substances in a circular economy,” *J. Environ. Manage.*, vol. 212, pp. 108–114, Apr. 2018, doi:

10.1016/j.jenvman.2018.02.014.

[41] M. Wahlström, T. Teittinen, T. Kaartinen, and van C. Liesbet, “Hazardous substances in construction products and materials: PARADE. Best practices for Pre-demolition Audits ensuring high quality RAw materials,” Esbo, Finland, 2019.

[42] Kretsloppsrådet, “Resurs och avfallshantering vid byggande och rivning,” 2019. .

[43] J. Powell, P. Jain, A. Bigger, and T. G. Townsend, “Development and Application of a Framework to Examine the Occurrence of Hazardous Components in Discarded Construction and Demolition Debris: Case Study of Asbestos-Containing Material and Lead-Based Paint,” *J. Hazardous, Toxic, Radioact. Waste*, vol. 19, no. 4, p. 05015001, 2015, doi: 10.1061/(asce)hz.2153-5515.0000266.

[44] D. Lundblad and M. Hult, “Farliga och miljöstörande material i hus guidebok om förekomst och hantering,” 2006.

[45] M. Govorko, L. Fritschi, and A. Reid, “Using a mobile phone app to identify and assess remaining stocks of in situ asbestos in australian residential settings,” *Int. J. Environ. Res. Public Health*, vol. 16, no. 24, 2019, doi: 10.3390/ijerph16244922.

[46] M. H. Govorko, L. Fritschi, and A. Reid, “Accuracy of a mobile app to identify suspect asbestos-containing material in Australian residential settings,” *J. Occup. Environ. Hyg.*, vol. 15, no. 8, pp. 598–606, 2018, doi: 10.1080/15459624.2018.1475743.

[47] M. H. Govorko, L. Fritschi, J. White, and A. Reid, “Identifying Asbestos-Containing Materials in Homes: Design and Development of the ACM Check Mobile Phone App,” *JMIR Form. Res.*, vol. 1, no. 1, p. e7, 2017, doi: 10.2196/formative.8370.

[48] F. Paglietti *et al.*, “Asbestos Risk: From Raw Material to Waste Management: The Italian Experience,” *Crit. Rev. Environ. Sci. Technol.*, vol. 42, no. 17, pp. 1781–1861, Sep. 2012, doi: 10.1080/10643389.2011.569875.

[49] S. G. B. Council, “Vad är Miljöbyggnad iDrift? - Sweden Green Building Council - Sweden Green Building Council.” https://www.sgbc.se/certifiering/miljobyggnad-idrift/vad-ar-miljobyggnad-idrift/ (accessed Oct. 29, 2021).

[50] L. A. Akanbi, A. O. Oyedele, L. O. Oyedele, and R. O. Salami, “Deep learning model for Demolition Waste Prediction in a circular economy,” *J. Clean. Prod.*, vol. 274, 2020, doi: 10.1016/j.jclepro.2020.122843.

[51] B. Soust-Verdaguer, C. Llatas, and A. García-Martínez, “Critical review of bim-based LCA method to buildings,” *Energy Build.*, vol. 136, pp. 110–120, 2017, doi: 10.1016/j.enbuild.2016.12.009.

[52] G. W. Cha, Y. C. Kim, H. J. Moon, and W. H. Hong, “New approach for forecasting demolition waste generation using chi-squared automatic interaction detection (CHAID) method,” *J. Clean. Prod.*, vol. 168, pp. 375–385, 2017, doi: 10.1016/j.jclepro.2017.09.025.

[53] M. Bilal *et al.*, “Big Data in the construction industry: A review of present status, opportunities, and future trends,” *Adv. Eng. Informatics*, vol. 30, no. 3, pp. 500–521, 2016, doi: 10.1016/j.aei.2016.07.001.

[54] G. W. Cha *et al.*, “Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 19, pp. 1–15, 2020, doi: 10.3390/ijerph17196997.

[55] Y. C. Kim and W. H. Hong, “Optimal management program for asbestos containing building materials to be available in the event of a disaster,” *Waste Manag.*, vol. 64, pp. 272–285, Jun. 2017, doi: 10.1016/j.wasman.2017.03.042.

[56] L. Fiumi, A. Campopiano, S. Casciardi, and D. Ramires, “Method validation for the identification of asbestos-cement roofing,” *Appl. Geomatics*, vol. 4, no. 1, pp. 55–64, 2012, doi: 10.1007/s12518-012-0078-0.

[57] E. Wilk, M. Krówczyńska, and P. Pabjanek, “Determinants influencing the amount of asbestos-cement roofing in Poland,” *Misc. Geogr.*, vol. 19, no. 3, pp. 82–86, 2015, doi: 10.1515/mgrsd-2015-0014.

[58] E. Wilk, M. Krówczyńska, and B. Zagajewski, “Modelling the spatial distribution of asbestos-cement products in Poland with the use of the random forest algorithm,” *Sustain.*, vol. 11, no. 16, 2019, doi: 10.3390/su11164355.

[59] M. Krówczyńska, E. Wilk, P. Pabjanek, and G. Olędzka, “Pleural mesothelioma in Poland: Spatial analysis of malignant mesothelioma prevalence in the period 1999-2013,” *Geospat. Health*, vol. 13, no. 2, pp. 314–321, 2018, doi: 10.4081/gh.2018.667.

[60] M. Krówczyńska, E. Raczko, N. Staniszewska, and E. Wilk, “Asbestos-cement roofing identification using remote sensing and convolutional neural networks (CNNs),” *Remote Sens.*, vol. 12, no. 3, pp. 1–16, 2020, doi: 10.3390/rs12030408.

[61] T. Mecharnia, L. C. Khelifa, N. Pernelle, and F. Hamdi, “An approach

toward a prediction of the presence of asbestos in buildings based on incomplete temporal descriptions of marketed products," in *K-CAP 2019 - Proceedings of the 10th International Conference on Knowledge Capture*, 2019, pp. 239–242, doi: 10.1145/3360901.3364428.

[62] G. Bonifazi, G. Capobianco, and S. Serranti, "Hyperspectral imaging and hierarchical PLS-DA applied to asbestos recognition in construction and demolition waste," *Appl. Sci.*, vol. 9, no. 21, pp. 1–15, 2019, doi: 10.3390/app9214587.

[63] G. Bonifazi, G. Capobianco, and S. Serranti, "Asbestos containing materials detection and classification by the use of hyperspectral imaging," *J. Hazard. Mater.*, vol. 344, pp. 981–993, 2018, doi: 10.1016/j.jhazmat.2017.11.056.

[64] O. Adedeji and Z. Wang, "Intelligent waste classification system using deep learning convolutional neural network," *Procedia Manuf.*, vol. 35, pp. 607–612, 2019, doi: 10.1016/j.promfg.2019.05.086.

[65] A. Rashidi, M. H. Sigari, M. Maghiar, and D. Citrin, "An analogy between various machine-learning techniques for detecting construction materials in digital images," *KSCE J. Civ. Eng.*, vol. 20, no. 4, pp. 1178–1188, 2016, doi: 10.1007/s12205-015-0726-0.

[66] P. Kuritcyn, K. Anding, E. Linß, and S. M. Latyev, "Increasing the safety in recycling of construction and demolition waste by using supervised machine learning," *J. Phys. Conf. Ser.*, vol. 588, no. 1, 2015, doi: 10.1088/1742-6596/588/1/012035.

[67] K. Anding, D. Garten, and E. Linß, "Application of intelligent image processing in the construction material industry," *Acta Imeko*, vol. 2, no. 1, p. 61, 2013, doi: 10.21014/acta_imeko.v2i1.100.

[68] K. Anding, E. Linß, H. Träger, M. Rückwardt, and A. Göpfert, "Optical identification of construction and demolition waste by using image processing and machine learning methods," *14th Jt. Int. IMEKO TC1, TC7, TC13 Symp. Intell. Qual. Meas. - Theory, Educ. Train. 2011, Held Conj. with 56th IWK Ilmenau Univ. Technol.*, pp. 126–132, 2011.

[69] T. Hong, Z. Wang, X. Luo, and W. Zhang, "State-of-the-art on research and applications of machine learning in the building life cycle," *Energy Build.*, vol. 212, p. 109831, 2020, doi: 10.1016/j.enbuild.2020.109831.

[70] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms, 3rd Edition | Wiley*. Hoboken, NJ, USA: Wiley-IEEE Press, 2019.

[71] M. Mangold, M. Österbring, H. Wallbaum, L. Thuvander, and P. Femenias, "Socio-economic impact of renovation and energy retrofitting of the

Gothenburg building stock," 2016, doi: 10.1016/j.enbuild.2016.04.033.

[72] F. Dalla Longa, B. Sweerts, and B. van der Zwaan, "Exploring the complex origins of energy poverty in The Netherlands with machine learning," *Energy Policy*, vol. 156, no. September 2020, p. 112373, 2021, doi: 10.1016/j.enpol.2021.112373.

[73] M. Mangold, M. Österbring, H. Wallbaum, L. Thuvander, and P. Femenias, "Socio-economic impact of renovation and energy retrofitting of the Gothenburg building stock," *Energy Build.*, vol. 123, pp. 41–49, Jul. 2016, doi: 10.1016/J.ENBUILD.2016.04.033.

[74] L. Fiumi, "Evaluation of MIVIS hyperspectral data for mapping covering materials," in *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, DFUA 2001*, 2001, pp. 324–327, doi: 10.1109/DFUA.2001.985906.

[75] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020, doi: 10.1109/ACCESS.2020.2976199.

[76] M. Brøgger and K. B. Wittchen, "Energy Performance Certificate Classifications Across Shifting Frameworks," *Procedia Eng.*, vol. 161, pp. 845–849, 2016, doi: 10.1016/j.proeng.2016.08.727.

[77] B. Hårsman, Z. Daghbashyan, and P. Chaudhary, "On the Quality and Impact of Residential Energy Performance Certificates," *CESIS Electron. Work. Pap. Ser.*, 2016, Accessed: Nov. 05, 2021. [Online]. Available: http://www.cesis.se.

[78] A. Simon, "Definition of validation levels and other related concepts," Luxembourg City, Luxembourg, 2013.

[79] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/S13748-016-0094-0.

[80] A. Darko, A. P. C. Chan, M. A. Adabre, D. J. Edwards, M. R. Hosseini, and E. E. Ameyaw, "Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities," *Autom. Constr.*, vol. 112, no. January, 2020, doi: 10.1016/j.autcon.2020.103081.

[81] S. Raschka and V. Mirjalili, *Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow*. Birmingham: Packt Publishing Ltd., 2017.

[82] Y. Chen, L. K. Norford, H. W. Samuelson, and A. Malkawi, "Optimal control of HVAC and window systems for natural ventilation through reinforcement learning," *Energy Build.*, vol. 169, pp. 195–205, 2018, doi:

10.1016/j.enbuild.2018.03.051.

[83] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, “Reinforcement learning for optimal control of low exergy buildings,” *Appl. Energy*, vol. 156, pp. 577–586, 2015, doi: 10.1016/j.apenergy.2015.07.050.

[84] M. Emami Javanmard, S. F. Ghaderi, and M. Hoseinzadeh, “Data mining with 12 machine learning algorithms for predict costs and carbon dioxide emission in integrated energy-water optimization model in buildings,” *Energy Convers. Manag.*, vol. 238, no. April, p. 114153, 2021, doi: 10.1016/j.enconman.2021.114153.

[85] X. J. Luo, L. O. Oyedele, A. O. Ajayi, and O. O. Akinade, “Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads,” *Sustain. Cities Soc.*, vol. 61, no. May, p. 102283, 2020, doi: 10.1016/j.scs.2020.102283.

[86] H. ur Rehman, T. Korvola, R. Abdurafikov, T. Laakko, A. Hasan, and F. Reda, “Data analysis of a monitored building using machine learning and optimization of integrated photovoltaic panel, battery and electric vehicles in a Central European climatic condition,” *Energy Convers. Manag.*, vol. 221, no. March, p. 113206, 2020, doi: 10.1016/j.enconman.2020.113206.

[87] T. Su, H. Li, and Y. An, “A BIM and machine learning integration framework for automated property valuation,” *J. Build. Eng.*, vol. 44, p. 102636, 2021, doi: 10.1016/j.jobe.2021.102636.

[88] C. Miller, “What’s in the box?! Towards explainable machine learning applied to non-residential building smart meter classification,” *Energy Build.*, vol. 199, pp. 523–536, 2019, doi: 10.1016/j.enbuild.2019.07.019.

[89] X. Gao and A. Malkawi, “A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm,” *Energy Build.*, vol. 84, pp. 607–616, 2014, doi: 10.1016/j.enbuild.2014.08.030.

[90] E. Wang, Z. Shen, and K. Grosskopf, “Benchmarking energy performance of building envelopes through a selective residual-clustering approach using high dimensional dataset,” *Energy Build.*, vol. 75, pp. 10–22, 2014, doi: 10.1016/j.enbuild.2013.12.055.

[91] G. Kropat *et al.*, “Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units,” *J. Environ. Radioact.*, vol. 147, pp. 51–62, 2015, doi: 10.1016/j.jenvrad.2015.05.006.

[92] F. Khayatian, L. Sarto, and G. Dall’O’, “Application of neural networks for evaluating energy performance certificates of residential buildings,”

Energy Build., vol. 125, pp. 45–54, 2016, doi: 10.1016/j.enbuild.2016.04.067.

[93] Y. He, J. Henze, and B. Sick, "Continuous learning of deep neural networks to improve forecasts for regional energy markets," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 12175–12182, 2020, doi: 10.1016/j.ifacol.2020.12.1017.

[94] "Towards the development of residential smart districts.pdf." .

[95] J. Wan *et al.*, "Deep learning for content-based image retrieval: A comprehensive study," *MM 2014 - Proc. 2014 ACM Conf. Multimed.*, pp. 157–166, Nov. 2014, doi: 10.1145/2647868.2654948.

[96] K. W. Brown, T. Minegishi, C. C. Cummiskey, M. A. Fragala, R. Hartman, and D. L. MacIntosh, "PCB remediation in schools: a review," *Environ. Sci. Pollut. Res.*, vol. 23, no. 3, pp. 1986–1997, 2016, doi: 10.1007/s11356-015-4689-y.

[97] D. Gough, S. Oliver, and J. Thomas, *An introduction to systematic reviews / David Gough, Sandy Oliver, James Thomas*. 2012.

[98] C. Chen, "Science Mapping: A Systematic Review of the Literature," *J. Data Inf. Sci.*, vol. 2, no. 2, pp. 1–40, 2017, doi: 10.1515/jdis-2017-0006.

[99] D. Moher *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, 2009, doi: 10.1371/journal.pmed.1000097.

[100] F. Community, "FME and GPS," 2020. https://community.safe.com/s/article/fme-and-gps (accessed Nov. 23, 2021).

[101] L. P. Fastighetsuttag, N.-D. I. G. O. C. H. Fastighetsavisering, A. Tj, and N. Gr, "Överföringsformatet i Fastighetsregistret," pp. 1–263, 2021.

[102] Boverket, "Sammanfattning av energideklaration." 2007.

[103] Boverket, "Energy performance certificate - Boverket," 2021. https://www.boverket.se/en/start/building-in-sweden/contractor/inspection-delivery/energy-performance-certificate/ (accessed Dec. 20, 2021).

[104] S. Sweden, "Storstadsområden Stor-Göteborg Stor-Stockholm Stor-Malmö." 2005.

[105] Hitta.se, "Karta med Tomtgränser i Sverige." https://www.hitta.se/kartan/tomter!~62.17408,14.93199,4z/tileLayer!l=1/realestate!a=1/tr!i=YxioN4ov (accessed Nov. 23, 2021).

[106] W. Mckinney, "Python for Data Analysis," Accessed: Dec. 20, 2021.

[Online]. Available: www.allitebooks.com.

[107] Matplotlib, “Matplotlib — Visualization with Python,” 2021. https://matplotlib.org/ (accessed Dec. 20, 2021).

[108] M. Waskom, “seaborn: statistical data visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/JOSS.03021.

[109] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, “What a p-value tells you about statistical significance,” *Am. Stat.*, vol. 73, no. sup1, pp. 1–19, Mar. 2019, doi: 10.1080/00031305.2019.1583913.

[110] J. Brownlee, “Feature Selection in Python with Scikit - Learn,” *Machine Learning Mastery*, 2020. https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/ (accessed Jul. 09, 2021).

[111] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Accessed: Jul. 13, 2021. [Online]. Available: https://github.com/catboost/catboost.

[112] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Jul. 09, 2021. [Online]. Available: https://github.com/slundberg/shap.

[113] R. L. Neitzel, S. K. Sayler, A. H. Demond, H. d’Arcy, D. H. Garabrant, and A. Franzblau, “Measurement of asbestos emissions associated with demolition of abandoned residential dwellings,” *Sci. Total Environ.*, vol. 722, p. 137891, 2020, doi: 10.1016/j.scitotenv.2020.137891.

[114] A. Campopiano, S. Casciardi, F. Fioravanti, and D. Ramires, “Airborne Asbestos Levels in School Buildings in Italy,” *J. Occup. Environ. Hyg.*, vol. 1, no. 4, pp. 256–261, 2004, doi: 10.1080/15459620490433771.

[115] R. F. Herrick, J. H. Stewart, and J. G. Allen, “Review of PCBs in US schools: a brief history, an estimate of the number of impacted schools, and an approach for evaluating indoor air samples,” *Environ. Sci. Pollut. Res.*, vol. 23, no. 3, pp. 1975–1985, 2016, doi: 10.1007/s11356-015-4574-8.

[116] D. L. MacIntosh *et al.*, “Mitigation of building-related polychlorinated biphenyls in indoor air of a school,” *Environ. Heal. A Glob. Access Sci. Source*, vol. 11, no. 1, pp. 1–10, Apr. 2012, doi: 10.1186/1476-069X-11-24.

[117] M. Kohler *et al.*, “Joint sealants: An overlooked diffuse source of polychlorinated biphenyls in buildings,” *Environ. Sci. Technol.*, vol. 39, no. 7, pp. 1967–1973, Apr. 2005, doi: 10.1021/es048632z.

[118] I. B. A. Secretariat, “Current Asbestos Bans,” 2019. http://ibasecretariat.org/alpha_ban_list.php (accessed Nov. 18, 2021).

[119] L. C. Oliver, N. L. Sprince, and R. Greene, “Asbestos-related disease in public school custodians,” *Am. J. Ind. Med.*, vol. 19, no. 3, pp. 303–316, 1991, doi: 10.1002/ajim.4700190305.

[120] J. Ameille *et al.*, “Pleural thickening: a comparison of oblique chest radiographs and high-resolution computed tomography in subjects exposed to low levels of asbestos pollution,” *Int. Arch. Occup. Environ. Health*, vol. 64, no. 8, pp. 545–548, Jan. 1993, doi: 10.1007/BF00517698.

[121] J. H. Lange, P. R. Lange, T. K. Reinhard, and K. W. Thomulka, “A study of personal and area airborne asbestos concentrations during asbestos abatement: A statistical evaluation of fibre concentration data,” *Ann. Occup. Hyg.*, vol. 40, no. 4, pp. 449–466, 1996, doi: 10.1016/0003-4878(95)00081-X.

[122] J. Peto, F. E. Matthews, J. T. Hodgson, and J. R. Jones, “Continuing increase in mesothelioma mortality in Britain,” *Lancet*, vol. 345, no. 8949, pp. 535–539, Mar. 1995, doi: 10.1016/S0140-6736(95)90462-X.

[123] V. Neumann, S. Günther, K. M. Müller, and M. Fischer, “Malignant mesothelioma - German mesothelioma register 1987-1999,” *Int. Arch. Occup. Environ. Health*, vol. 74, no. 6, pp. 383–395, 2001, doi: 10.1007/s004200100240.

[124] A. Marinaccio *et al.*, “Analysis of latency time and its determinants in asbestos related malignant mesothelioma cases of the Italian register,” *Eur. J. Cancer*, vol. 43, no. 18, pp. 2722–2728, Dec. 2007, doi: 10.1016/j.ejca.2007.09.018.

[125] S. Y. Kim, Y. C. Kim, Y. Kim, and W. H. Hong, “Predicting the mortality from asbestos-related diseases based on the amount of asbestos used and the effects of slate buildings in Korea,” *Sci. Total Environ.*, vol. 542, pp. 1–11, Jan. 2016, doi: 10.1016/j.scitotenv.2015.10.115.

[126] V. Bourdès, P. Boffetta, and P. Pisani, “Environmental exposure to asbestos and risk of pleural mesothelioma: Review and meta-analysis - Environmental exposure to asbestos and mesothelioma,” *Eur. J. Epidemiol.*, vol. 16, no. 5, pp. 411–417, 2000, doi: 10.1023/A:1007691003600.

[127] B. Armstrong and T. Driscoll, “Mesothelioma in Australia: Cresting the third wave,” *Public Heal. Res. Pract.*, vol. 26, no. 2, Apr. 2016, doi: 10.17061/phrp2621614.

[128] F. Frassy *et al.*, “Mapping asbestos-cement roofing with hyperspectral remote sensing over a large mountain region of the Italian western alps,” *Sensors (Switzerland)*, vol. 14, no. 9, pp. 15900–15913, 2014, doi: 10.3390/s140915900.

[129] M. Kohler, M. Zennegg, and R. Waeber, “Coplanar polychlorinated biphenyls (PCB) in indoor air,” *Environ. Sci. Technol.*, vol. 36, no. 22, pp. 4735–4740, Nov. 2002, doi: 10.1021/es025622u.

[130] E. Priha, S. Hellman, and J. Sorvari, “PCB contamination from polysulphide sealants in residential areas - Exposure and risk assessment,” *Chemosphere*, vol. 59, no. 4, pp. 537–543, 2005, doi: 10.1016/j.chemosphere.2005.01.010.

[131] R. A. Rudel, L. M. Seryak, and J. G. Brody, “PCB-containing wood floor finish is a likely source of elevated PCBs in residents’ blood, household air and dust: A case study of exposure,” *Environ. Heal. A Glob. Access Sci. Source*, vol. 7, no. 1, p. 2, Dec. 2008, doi: 10.1186/1476-069X-7-2.

[132] M. Robson, L. Melymuk, S. A. Csiszar, A. Giang, M. L. Diamond, and P. A. Helm, “Continuing sources of PCBs: The significance of building sealants,” *Environ. Int.*, vol. 36, no. 6, pp. 506–513, 2010, doi: 10.1016/j.envint.2010.03.009.

[133] T. Schettgen, A. Alt, D. Preim, D. Keller, and T. Kraus, “Biological monitoring of indoor-exposure to dioxin-like and non-dioxin-like polychlorinated biphenyls (PCB) in a public building,” *Toxicol. Lett.*, vol. 213, no. 1, pp. 116–121, Aug. 2012, doi: 10.1016/j.toxlet.2011.06.005.

[134] D. L. MacIntosh *et al.*, “Mitigation of building-related polychlorinated biphenyls in indoor air of a school,” *Environ. Heal. A Glob. Access Sci. Source*, vol. 11, no. 1, p. 24, Dec. 2012, doi: 10.1186/1476-069X-11-24.

[135] H. W. Meyer *et al.*, “Plasma polychlorinated biphenyls in residents of 91 PCB-contaminated and 108 non-contaminated dwellings-An exposure study,” *Int. J. Hyg. Environ. Health*, vol. 216, no. 6, pp. 755–762, Nov. 2013, doi: 10.1016/j.ijheh.2013.02.008.

[136] G. M. Lehmann, K. Christensen, M. Maddaloni, and L. J. Phillips, “Evaluating health risks from inhaled polychlorinated biphenyls: Research needs for addressing uncertainty,” *Environ. Health Perspect.*, vol. 123, no. 2, pp. 109–113, 2015, doi: 10.1289/ehp.1408564.

[137] X. Liu *et al.*, “Laboratory study of PCB transport from primary sources to building materials,” *Indoor Built Environ.*, vol. 25, no. 4, pp. 635–650, Jul. 2016, doi: 10.1177/1420326X15623355.

[138] D. Abriha, Z. Kovács, S. Ninsawat, L. Bertalan, B. Balázs, and S. Szabó, “Identification of roofing materials with discriminant function analyand random forest classifiers on pan-sharpened worldview-2 imagery – A comparison,” *Hungarian Geogr. Bull.*, vol. 67, no. 4, pp. 375–392, 2018, doi: 10.15201/hungeobull.67.4.6.

[139] M. B. A. Gibril, H. Z. M. Shafri, and A. Hamedianfar, “New semi-automated mapping of asbestos cement roofs using rule-based object-based image analysis and Taguchi optimization technique from WorldView-2 images,” *Int. J. Remote Sens.*, vol. 38, no. 2, pp. 467–491, 2017, doi: 10.1080/01431161.2016.1266109.

[140] Göteborgs Stad, “Miljöinventering inför rivning,” 2018.

[141] B. Kretsloppsrådet, “Resurs- och avfallsriktlinjer vid byggande och rivning,” 2017.

[142] N. Samuelsson, “Förändra varsamt vägledning vid ombyggnader av rekordårens bebyggelse,” *Riksantikvarieämbetet*, 2004. .

[143] J. Brownlee, “Cost-Sensitive Learning for Imbalanced Classification,” 2020. https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/ (accessed Nov. 27, 2021).

[144] S. Sweden, “Just over 5 million dwellings in Sweden,” 2021. https://www.scb.se/en/finding-statistics/statistics-by-subject-area/housing-construction-and-building/housing-construction-and-conversion/dwelling-stock/pong/statistical-news/dwelling-stock-2020-12-31/ (accessed Nov. 27, 2021).

[145] M. Bengtsson, “Energideklarationer i Sverige,” 2008.

[146] L. Engberg, “Miljöutredning för Göteborgs Stads Lokalförvaltning,” 2013.

Appendix A

Table A1. The validation metric describes the interpretation scores for each of the ten observed properties. The summarized table below is the aggregated results from ten pairs of interpretation scores. In each pair, the aligned interpretation was marked 0, the opposite denoted as 1. Then the results from each pair were further compared to identify the disagreement. The aggregated results show the interpretation scores vary significantly between properties, which may be due to property complexity and information clearness in different inventory types. Also, building information and demolition-related information were interpreted differently. Higher agreements on asbestos detection, construction year, asbestos-containing pipe insulation detection, asbestos-containing carpet glue detection, PCB detection, and mercury-containing light tube detection were reached. The outcome from the exercise led to a consensus on a rigorous routine for interpretation of information in audits documentation and reevaluation of data collection and compilation.

Property*	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Total
Building information											
Construction year	0	1	0	0	0	0	1	0	1	1	**4**
Renovation year	1	1	1	0	1	1	1	1	1	1	**9**
Number of floors	1	0	1	0	1	1	1	0	1	1	**7**
Floor area	0	1	1	1	0	1	1	1	1	1	**8**
Demolition-related information											
Investigation scope	0	0	1	0	1	1	1	0	1	1	**6**
Decontamination history	1	0	0	0	1	1	1	1	1	1	**7**
Asbestos	0	0	0	0	0	0	0	0	0	0	**0**
Pipe insulation	0	0	1	0	0	1	1	0	0	1	**4**
Valves	0	0	1	0	1	1	1	1	1	1	**7**
Door/windows insulation	0	0	1	0	1	1	1	1	0	1	**6**
Cement panel	0	0	1	1	1	1	1	1	1	0	**7**
Tile/clinker	0	0	1	0	1	1	1	1	1	1	**7**
Carpet glue	0	0	1	0	0	1	1	1	0	1	**5**
Floor mat	1	0	1	0	1	1	1	1	0	1	**7**
Switchboard	1	0	1	0	1	1	1	1	1	1	**8**
Joint	1	0	1	0	1	1	1	1	1	1	**8**
PCB	1	0	1	1	0	0	0	1	0	1	**5**
Joint/sealant	0	0	1	0	1	0	1	1	1	1	**6**
Sealed double windows	0	0	1	0	1	1	1	1	1	1	**7**
Capacitors	0	0	1	0	1	1	1	1	0	1	**6**
Acrylic flooring	0	0	1	0	1	1	1	1	1	1	**7**
Door closer	1	0	1	0	0	1	1	1	1	1	**7**
Oil in cable	1	0	1	0	1	1	1	1	1	1	**8**
CFC	1	1	1	1	1	1	1	0	1	1	**9**
Fridge/freezer	1	0	1	0	1	1	1	1	1	1	**8**
Building insulation	1	0	1	0	1	1	1	1	1	1	**8**
Cooling unit	1	0	1	0	1	1	1	1	1	1	**8**
Rolling gate	1	0	1	0	1	1	1	0	1	1	**7**
Mercury	0	1	1	1	0	1	0	0	1	1	**6**
Lighting tube	0	0	1	0	0	1	1	0	1	1	**5**
Relay/switch	0	0	1	0	1	1	1	1	1	1	**7**
Level sensor	0	0	1	0	1	1	1	1	1	1	**7**
Thermometer	0	0	1	0	1	1	1	1	1	1	**7**
Thermostat	0	0	1	0	1	1	1	1	1	1	**7**
Water lock	0	0	1	0	1	1	1	1	1	1	**7**
Low energy lamp	1	0	1	0	1	1	1	1	1	1	**8**
Doorbell	1	0	1	0	1	1	1	1	1	1	**8**
Total	**21**	**10**	**39**	**10**	**33**	**39**	**40**	**33**	**34**	**39**	**298**

* Property 1: Warehouse/Industrial building/Office from protocols and demolition plans.
Property 2: Single-family house from control plans.
Property 3: Multifamily house from demolition plans.
Property 4: Complementary building from demolition plans.
Property 5: School from reports.
Property 6: Multifamily house from reports and demolition plans.
Property 7: Office/Commercial building from reports.
Property 8: Industrial building from reports.
Property 9: Industrial building from reports and demolition plans.
Property 10: Single-family house from demolition plans.

Appendix B

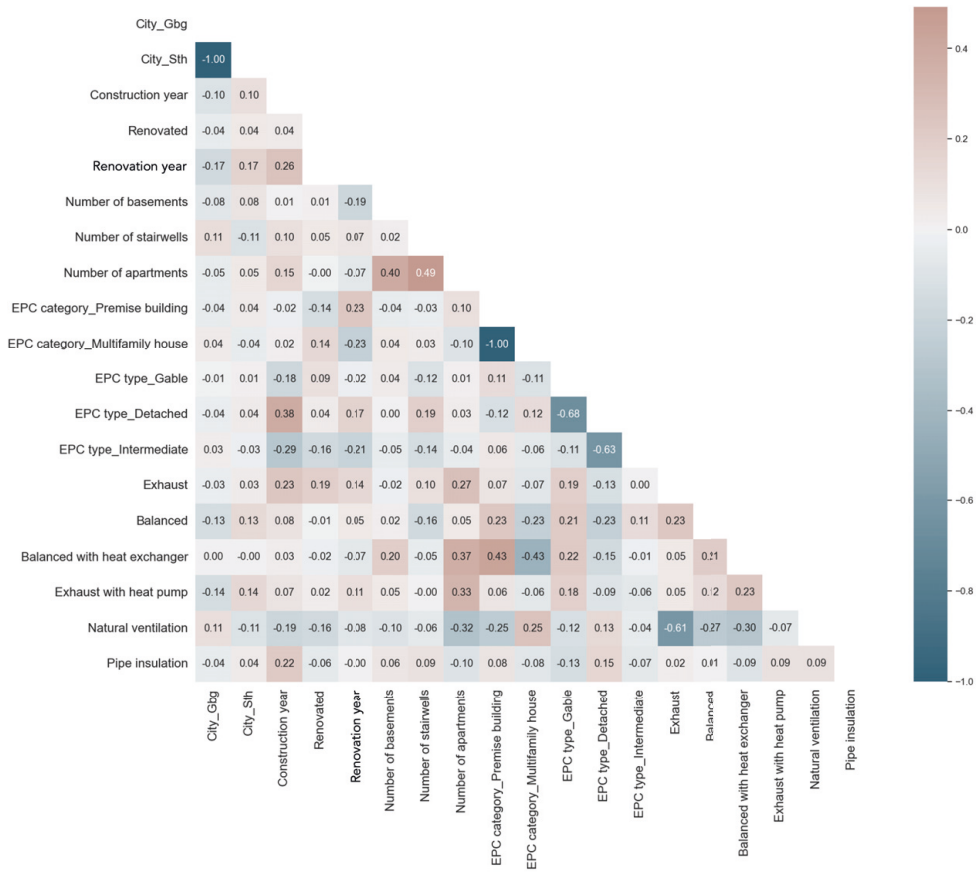


Figure (A2.1) Correlation plot for asbestos pipe insulation in multifamily houses.

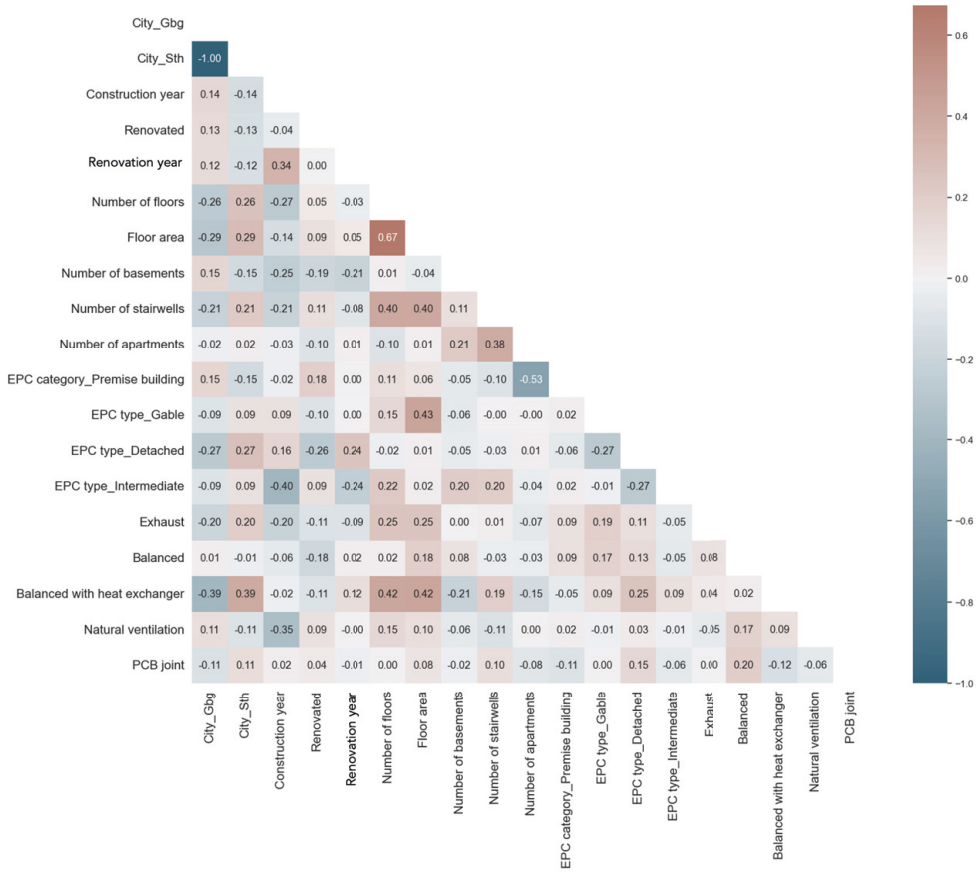


Figure (A2.2) Correlation plot for PCB joints in schools.

Figure A2. Pairwise correlation plot with Pearson's correlation coefficients (*r*). Directions of the correlation were shown in red (positive), white (neutral), and blue (negative) colors. No strong correlation ($|r| \geq 0.7$) was found for asbestos pipe insulation and PCB joints.

Appendix C

Table A3. Stepwise logistic regression shows correlations (r) their significance (*p*) between the target variable asbestos-containing pipe insulation and the other predictive variables in multifamily houses. The categorical variables were transformed to dummy variables using one-hot-encoding, then removed one of the dummy variables in each category as comparative values for regression analysis.

	Model 1	Model 2	Model 3
Construction year	0.03**	0.03*	0.02
Renovation year	-0.01	-0.02	-0.02
Number of floors	-0.11	-0.13	-0.11
Floor area	0.00	0.00	0.00
Number of basements	0.12	0.15	0.19
Number of stairwells	0.09	0.07	0.07
Number of apartments	-0.01*	-0.02	-0.02*
Exhaust		0.66	0.78
Balanced		0.10	0.12
Balanced with heat exchanger		-0.51	-0.85
Natural ventilation		0.62	0.76
Renovated		-0.46	-0.24
City_Gothenburg			-0.26
EPC building category_Multifamily house			-1.48
EPC building type_Gable			-0.62
EPC building type_Intermediate			-0.12
Constant	-24.21	-14.81	-0.20
No. observations	139	139	139

Table A4. Stepwise logistic regression shows correlations (*r*) their significance (*p*) between the target variable PCB-containing joint or sealant and the other predictive variables in school buildings.

	Model 1	Model 2	Model 3
Construction year	0.01	0.01	0.01
Renovation year	-0.00	0.00	0.01
Number of floors	-0.23	-0.10	-0.03
Floor area	0.00	0.00	0.00
Number of basements	0.08	-0.36	-0.10
Number of stairwells	0.31	0.40	0.50
Number of apartments	-0.19	-0.39	-0.19
Exhaust		-0.25	-0.12
Balanced		0.93	0.99
Balanced with heat exchanger		-1.21*	-1.56**
Natural ventilation		-24.61	-24.11
Renovated		-0.10	0.13
City_Gothenburg			-0.63
EPC building category_Multifamily house			-3.00*
EPC building type_Gavel			-0.85
EPC building type_Mellanliggande			-25.20
Constant	-6.42	-16.45	-27.97
No. observations	109	109	109