LUND UNIVERSITY

Data-driven Approaches for Predicting Hazardous Substances in the Building Stock

Wu, Pei-Yu

2024

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

Data-driven Approaches for Predicting Hazardous Substances in the Building Stock

PEI-YU WU | FACULTY OF ENGINEERING | LUND UNIVERSITY

Data-driven Approaches for Predicting Hazardous Substances in the Building Stock

Pei-Yu Wu

LUND UNIVERSITY

DOCTORAL THESIS

Doctoral thesis for the degree of Doctor of Philosophy (PhD) at the Faculty of Engineering at Lund University to be publicly defended on the 19th of January, 2024 at 13.15 in V:A Hall (V-huset), Department of Building and Environmental Technology, John Ericssons väg 1, Lund.

Faculty opponent
Dan Engström
Adjunct professor at the Department of Science and Technology
Division of Communications and Transport Systems
Linköping University
Linköping, Sweden

Data-driven Approaches for Predicting Hazardous Substances in the Building Stock

Pei-Yu Wu

LUND
UNIVERSITY

Cover photo by Pei-Yu Wu

Faculty of Engineering
Department of Building and Environmental Technology

ISBN (print) 978-91-88722-82-9
ISBN (e-version) 978-91-88722-83-6
ISSN 0349-4950

Printed in Sweden by Media-Tryck, Lund University
Lund 2023

We need to radically improve our relationship with nature, and that requires rethinking decisions and reorganising many processes.

Leila Nachawati and Maja Romano (2023)

Table of Contents

Abstract

The presence of hazardous substances in buildings introduces significant challenges to project scheduling, budget estimation, and occupant and worker safety in renovation and demolition activities. In Swedish renovation and demolition projects, allocating approximately 15% of the budget for unforeseen events and encountering of unexpected substances has become a standard practice. However, the actual costs for abatement and decontamination often exceed these estimates. Given the urgency to renovate aging buildings to current standards and the increasing focus on circular construction for material reuse and recycling, developing predictive tools for hazardous substances in buildings is crucial. Leveraging environmental data from pre-demolition audits, indoor radon measurements, and advanced algorithms offers new avenues for predicting hazardous substances. This thesis investigates the application of data-driven methods in predicting and interpreting patterns of in situ hazardous substances in existing buildings.

A comprehensive literature review establishes the thesis foundation, highlighting the lack of cost-effective methods for predicting hazardous substances at the building level. To bridge this gap, regional hazardous material and national indoor radon databases were compiled, with rigorous data quality and quantity assessments. Inspection records and radioactive substance measurements were digitized and integrated with building registers, enabling detailed analysis of detection rates across various building types and municipalities. This thesis advances the use of such data in statistical, machine learning, and neural network models to predict the presence of hazardous substances. The performance of these models was evaluated for their ability to estimate the probability and geospatial distribution of buildings likely containing substances such as polychlorinated biphenyl, asbestos, radioactive concrete, and high indoor radon levels.

The predictive models' outcomes offer insights into the occurrence and frequency of hazardous substances, enabling the implementation of risk-based pre-demolition inspections and circular construction management. The proposed method is adaptable, scalable and suitable for inventorying hazardous substances in regional and national building stocks. Continuous data integration from various regions, while maintaining representativeness of the Swedish building stock improved the models' generalizability and robustness. These predictions can guide policy design, helping authorities and municipalities screen and remediate contaminated buildings. They also assist in assessing uncertainties of hazardous

substances in building maintenance, renovation, demolition, and supporting building owners and contractors in safe and compliant practices.

Keywords: Hazardous material, Pre-demolition audit, Building stock, Machine learning, Risk assessment, Circular construction

Sammanfattning

Förekomst av farliga ämnen i byggnader medför betydande osäkerheter i projektplanering och budgetuppskattning och innebär hälso och säkerhetsrisk för de boende och för arbetstagare som är involverade i renoverings- och rivningsaktiviteter. Det är praxis inom svenska byggprojekt att avsätta ungefär 15% av budgeten för att hantera oväntade risker av farliga ämnen och oförutsedda händelser. De faktiska kostnaderna för sanering och dekontaminering överstiger dock ofta dessa uppskattningar. Givet behovet av att renovera åldrande byggnader för att uppfylla dagens byggnadsstandard och det ökande fokuset på cirkulärt byggande för att främja återanvändning och återvinning av material, är utveckling av nya verktyg för att förutsäga närvaron av farliga byggsubstanser av yttersta vikt. Tillgängligheten av miljödata från inventeringar före rivning och mätningar av inomhusradon, tillsammans med tillgången till avancerade algoritmer, erbjuder nya möjligheter att förutsäga förekomst av farliga ämnen. Avhandlingen syftar till att undersöka potentialen att använda datadrivna metoder för att förutsäga och tolka mönster av förekomst av farliga ämnen i det befintliga byggnadsbeståndet.

En omfattande litteraturgenomgång lägger grunden för denna avhandling och belyser bristen på kostnadseffektiva metoder för att förutsäga förekomsten av olika farliga ämnen på byggnadsnivå. För att adressera dessa kunskapsluckor skapades en regional databas för farliga ämnen och en nationell databas för inomhusradon, vilka genomgick rigorösa utvärderingar av datakvalitet och kvantitet. Inventeringsprotokoll och mätningar av radioaktiva ämnen digitaliserades systematiskt och integrerades med byggnadsregistret, vilket möjliggjorde mer detaljerade analyser av detektionsfrekvenser i olika byggnadstyper och kommuner. Avhandlingen banar vägen för användningen av dessa data för att träna statistiska modeller, maskininlärningsmodeller och neurala nätverksmodeller för att förutsäga förekomst av farliga ämnen. Modeller utvärderades, tolkades och användes därefter för att uppskatta sannolikheten för geografiska fördelningar av byggnader som potentiellt innehåller farliga ämnen så som polyklorerade bifenyler, asbestmaterial, radioaktiv betong och höga inomhusradonnivåer.

De prediktiva modellernas utfall ger insikter i förekomsten av farliga ämnen och hjälper till att möjliggöra genomförandet av riskbaserade inspektioner före rivning och cirkulärt byggande. Den föreslagna metoden är anpassningsbar, skalbar och lämplig för inventering av farliga ämnen i regionala och nationella byggnadsbestånd. Kontinuerlig dataintegration från olika regioner, med bibehållen representativitet för det svenska byggnadsbeståndet, förbättrade modellernas generaliserbarhet och robusthet. Dessa förutsägelser kan vägleda policyutformning, hjälpa myndigheter, kommuner och fastighetsägare att screena och åtgärda byggnader där farliga ämnen förekommer. De

hjälper också till att bedöma osäkerheter om farliga ämnen vid byggnadsunderhåll, renovering, rivning och hjälper byggnadsägare och entreprenörer att välja säkra metoder i enlighet med rådande regelverk.

Nyckelord: Farliga ämnen, Miljöinventering, Byggnadsbestånd, Maskininlärning, Riskbedömning; Cirkulärt byggande

摘要

建築物中的有害物質對工程專案進度和經費估算造成重大的不確定性，並且引發對居民和執行建築翻新、拆除工作的工人的安全疑慮。在瑞典的建築改造和拆除工程中，標準做法是將約 15%的預算分配於因應不可預見的事件和處置有害物質。然而，實際上清除和去污的成本經常超過這些估算。鑑於迫切翻新老舊建築以滿足現代建築標準需求，並且日益強調落實循環建築實踐，以促進材料再使用和回收的趨勢，因此需要開發新的方法來預測建築物中可能出現的有害物質。利用長年累積的建築物拆前有害廢棄物的清查紀錄和室內氡氣測量記錄等環境數據，結合先進的演算法，為研究建築物中的有害物質提供了新的可能性。本論文旨在探討將資料導向的方法應用於預測和解釋既有建築群中有害物質的潛力。

為了奠定論文的理論基礎，本論文進行了全面的文獻探討，研究結果闡明了缺乏符合經濟效益的方法來預測建築中的有害物質。為了解決研究領域的不足，本論文創建了一個區域性有害物質資料庫和一個全國性室內氡氣的資料庫，並對其數據數量和質量進行了評估。藉由將建築物中的有害物質清查記錄和放射性物質測量數位化，並與建築資料庫相整合，從而更詳細地分析有害物質在各種建築類型和市鎮的出現模式。本論文進一步利用了這些數據來訓練統計模型、機器學習模型和神經網路模型，用以預測建築物中有害物質的存在。隨後，對這些預測模型的性能進行評估，估算其預測建築物中可能包含多氯聯苯材料、石棉材料、放射性混凝土和高濃度室內氡氣水平等物質的機率和地理空間分佈的效能。

預測模型的結果闡明建築物中有害物質的出現頻率，增加對有害物質出現模式的理解，並輔助實施風險管理為基準的有害物質清查，制定環境友善和安全的廢棄物管理計畫。本論文所提出的預測方法具有可適應性和可擴展性，適用於預測區域性和全國性建築群中的有害物質。透過持續整合來自不同地區的建築環境數據，並確保資料樣本對瑞典建築群的代表性，提高模型的通用性和穩健性。本論文的研究成果可以引導相關政府部門進行資料導向的決策，設計有效的政策工具篩選和清理受污染建築物，並協助評估建築維護、翻新和拆除的有害物質風險，從而確保建築業主和承包商採取安全和合規的工程實踐。

關鍵字: 有害物質、廢棄物清查、建築群、機器學習、風險評估、循環建築實踐

Popular Science Summary

Addressing hazardous materials is a critical step for the construction industry to grapple with the challenges of sustainability and circular practices. Over the past century, hazardous materials have been used in construction worldwide unintendedly. With a large number of existing buildings approaching the end of their service life, renovations and selected demolitions are required to meet modern standards and ensure comfortable living environments. However, these efforts are often hindered by the high risk of encountering hazardous materials unexpectedly during renovation and demolition, leading to project delays, cost overruns, and safety concerns for the reuse and recycling of reclaimed building components in their next lifecycles.

To address this risk, many European countries have made pre-demolition audits mandatory or strongly recommended as a means to characterize and quantify in situ hazardous materials to ensure safe construction and demolition waste management. Nevertheless, the research in the field of hazardous materials in circular building practices has often been fragmented, existing in "disciplinary silos" within different domains of environmental science, public health, or engineering. Regulations also sometimes clash, with differences between chemical regulation and circular economy policies. So far, no holistic studies have tackled diverse hazardous materials in the context of the building industry from an overarching perspective.

This thesis sought to bridge these knowledge gaps by proposing a comprehensive method for predicting in situ hazardous substances and benchmarking the performance of different predictive modeling approaches. It utilized historical environmental inventories and indoor radon measurements, creating digital datasets on a building-by-building basis for predictive modeling and pattern identification. Machine learning and statistical pipelines were developed to enable effective data analysis, model training, evaluation, and prediction for hazardous materials and radioactive substances in existing buildings. A data-driven approach that offers decision support for stakeholders was developed, enabling cost-efficient in situ hazardous substance assessment for both individual buildings and large-scale building stock. The prediction outcomes, with their probability distributions, are expected to guide risk-based inspections in existing buildings likely to have contamination. This holistic approach is a crucial building block for a future where construction, renovation, reconstruction, and demolition is both resource-efficient and environmentally responsible.

List of Publications

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

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), (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(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
Building and Environment 213(108894), (2022)

IV *Machine learning models for the prediction of polychlorinated biphenyls and asbestos materials in buildings*
P-Y. Wu, C. Sandels, T. Johansson, M. Mangold, and K. Mjörnell
Resources, Conservation, and Recycling 199(107253), (2023)

V *Estimating the Probability Distributions of Radioactive Concrete in the Building Stock Using Bayesian Networks*
P-Y. Wu, T. Johansson, M. Mangold, C. Sandels, and K. Mjörnell
Expert Systems with Applications 222(119812), (2023)

VI *Indoor Radon Interval Prediction in the Swedish Building Stock Using Machine Learning*
P-Y. Wu, T. Johansson, C. Sandels, M. Mangold, and K. Mjörnell
Building and Environment 245(110879), (2023)

Conf I *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)

Conf II *Evaluating the Indoor Radon Concentrations in the Swedish Building Stock Using Statistical and Machine Learning*
P-Y. Wu, T. Johansson, M. Mangold, C. Sandels, and K. Mjörnell
J. Phys.:Conf. Ser. (2023) (published)

Other related publications by the author:

Licentiate thesis *Predicting hazardous materials in the Swedish building stock using data mining*
P-Y. Wu
Media-Tryck, Lund University (2022)

Modeling Artificial Neural Networks to Predict Asbestos-containing Materials in Residential Buildings
P-Y. Wu, M. Mangold, C. Sandels, T. Johansson, and K. Mjörnell
IOP Conf. Ser.: Earth Environ. Sci. 1122(012050) (2022)

Contribution of the Publications

Paper I
I and the second author conducted a literature search. Then I collected data and conducted formal analysis, drafted and revised the manuscript. The fourth author guided me in addressing comments from peer review and disseminated the publication. All co-authors participated in the results discussion and manuscript review.

Paper II
I collected data and conducted a formal analysis, drafted and revised the manuscript. The third author established contact for data assembling and assisted with the paper layout. The fourth author contributed to method development and publication dissemination. The fifth author supplied a building database. All co-authors participated in cross-validation workshops, result discussions, and manuscript review.

Paper III
I collected data, developed methods, conducted formal analysis, drafted and revised the manuscript. The second and the fourth authors conceived the idea and contributed to method development. The fifth author supplied a building database. The fourth author contributed to the publication dissemination. All co-authors participated in the results discussion and manuscript review.

Paper IV
I conceived the idea, collected data, developed methods, conducted formal analysis, drafted and revised the manuscript. The second and the third authors assisted with data curation and result discussion. All co-authors contributed to the manuscript review.

Paper V
I conceived the idea, collected data, developed methods, conducted formal analysis, drafted and revised the manuscript. The second author curated the database. All co-authors participated in the results discussion and manuscript review.

Paper VI
I conceived the idea, collected data, developed methods, conducted formal analysis, drafted and revised the manuscript. The second author investigated and curated the database. The third author contributed to method development. All co-authors participated in the results discussion and manuscript review.

Conf I
I collected data, conducted a formal analysis, drafted and revised the manuscript. All co-authors participated in the results discussion and manuscript review.

Conf II
I conceived the idea, collected data, developed methods, conducted formal analysis, drafted and revised the manuscript. The second author investigated and curated the database. All co-authors participated in the results discussion and manuscript review.

List of Figures

List of Tables

Acknowledgments

This doctoral thesis stands as a testament to the support and collaboration of many, whose shared passion for research and contributions that have been invaluable. I am deeply grateful to my supervisors, Kristina Mjörnell, Mikael Mangold, and Claes Sandels, for their insightful guidance. Their unique "patchwork" of expertise and their trust in me have been crucial in navigating the challenges of research, encouraging me to think independently. Profuse appreciation go to my colleagues at RISE, Tim Johansson and Mikael Theorin, for their fully support in data curation and domain knowledge, which have opened new directions and added depth to the research.

The progression of data collection would not have been possible without the dedicated efforts of my students Frida Palstam, Birgitta Gunér, and Olivia Heuts, along with generous data provision from the City Archives of Gothenburg, Stockholm, Kiruna, the Environmental Administration of Malmö, the Swedish National Board of Housing, Building and Planning, the Swedish Land Survey, and the Geological Survey of Sweden, has been fundamental.

I extend my heartfelt appreciation to the industrial representatives for their valuable inputs and to RISE for granting me the academic freedom to pursue this work. My colleagues at RISE and LTH Building Physics and Building Service have enriched my experience with stimulating intellectual discussions. Being affiliated with Lund University has been an honor, and the opportunities for international outreach, including Erasmus+ and collaborations with Chile, have been immensely rewarding.

A special note of gratitude to Catherine De Wolf at ETH Zürich's Chair of Circular Engineering in Architecture for the unforgettable research exchange experience and for including me in her course. My opponents and examining committee for reading and review of my licentiate and doctoral thesis, the anonymous editors and reviewers, and my funders – the Swedish Foundation for Strategic Research, the European Union, Maj och Hilding Brosenius Research Foundation, and the Swedish Energy Agency – all deserve my sincere thanks.

Finally, this thesis is dedicated to my family: dad, mom, my sisters, my babysitters, my German mom, and my beloved husband Amir. Their unwavering support, love, and endless encouragement have been the cornerstone of my journey.

Pei-Yu Wu

Gothenburg, November 2023

Definitions and Acronyms

Asbestos-containing materials (ACM)

Asbestos has different forms in construction products, such as chrysotile, amosite, and crocidolite. They can cause asbestosis, mesothelioma, and lung cancer at high levels of exposure.

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.

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

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 and 1.0.

Building Information Model (BIM)

BIM is a digital representation of the physical and functional characteristics of a facility that manages information on a construction project throughout its life cycle.

Bayesian quantile regression

Bayesian quantile regression is a statistical approach combining quantile regression with Bayesian inference that models the conditional distribution of a dependent variable, allowing for the direct estimation of different quantiles while incorporating prior knowledge and handling uncertainty.

Bayesian Information Criteria (BIC)

BIC describes how well a model captures the underlying structure of the data and is used for model selection. It regulates model complexity by introducing a penalty under the maximum likelihood estimation.

$\mathrm{BIC} = \log(n)k - 2\log(\hat{L})$

n = the number of data points
k = the number of free parameters to be estimated
$\hat{L}$ = the maximized value of the likelihood function of the model

Circular Economy (CE)

CE is a model of production and consumption that applies principles of sharing, leasing, reusing, repairing, refurbishing, and recycling existing materials and products to extend the life cycle of products.

Cohen's kappa

Cohen's kappa coefficient is a statistic that is used to measure inter-rater or intra-rater reliability for qualitative items. It measures the agreement between two raters who each classify N items into C mutually exclusive categories.

Data mining

Data mining is an analytical technology that extracts and analyzes underlying relationships from large-amount and multi-attribute information.

Data corruption

Data corruption refers to errors in computer data due to unintended changes to the original data that may occur during the storage, writing, reading, transmission, or processing of data.

Deep neural network (DNN)

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

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based clustering algorithm that finds core samples of high density and expands clusters from vector array or distance matrix.

Energy Performance Certificates (EPC)

EPC is a rating scheme that provides information on the building's energy consumption produced by an independent certified energy expert and the certificate is valid for ten years.

Empirical Distribution Matching (EDM)

EDM is a bias-correction method and is implemented by obtaining the inverse empirical cumulative distribution function for the observed values and the machine learning estimate at the distribution scale.

F1

F1 is a harmonized mean of Precision (PRE) and Recall (REC) and works well for imbalanced data. It is a scale variable with a range between 0-1.0.

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

Micro-F1: Calculate metrics globally by counting the total true positives, false negatives, and false positives.

Macro-F1: Calculate metrics for each label and find their unweighted mean. This does not take label imbalance into account.

Weighted-F1: Calculate metrics for each label, and find their average weighted by the number of true instances for each label.

Geographical Information System (GIS)

GIS is a computer system and software that stores, manages, analyzes, and maps geographically reference information.

Gaussian diffusion model

Gaussian diffusion models use Gaussian processes to simulate or describe the diffusion of particles, heat, information, or other quantities over time and space.

Hierarchical clustering

Hierarchical clustering is an unsupervised learning method for cluster analysis that builds a hierarchical structure of data in a dendrogram (hierarchical tree) to reveal relationships among clusters.

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.

Kernal regression

Kernel regression is a type of non-parametric statistical technique used for estimating the conditional expectation of a random variable. It estimates relationships between variables when the underlying relationship is unknown or complex, which is particularly effective in smoothing and curve-fitting applications.

K-means clustering

K-means clustering is an unsupervised learning method for cluster analysis that partitions n datapoints into *K* clusters where the sum of the squared distances between the objects and their assigned cluster mean is minimized.

Life Cycle Analysis (LCA)

LCA is a methodology for assessing the environmental impacts such as emissions and resource use throughout the life cycle of a product or a service.

Material Flow Analysis (MFA)/ Material Stock Analysis (MSA)

MFA and MSA are scalable environmental accounting approaches for quantifying specific or multiple materials at various geographical and institutional scales.

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.

Missing Completely at Random (MCAR)

The missingness of data is independent both of observable variables and unobservable parameters of interest. It occurs entirely at random and the analysis performed is unbiased.

Missing at Random (MAR)

The missingness of data can be fully accounted for by variables where there is complete information. To prevent induced parameter bias in analysis, the parameter can be estimated asymptotically with Full Information on Maximum Likelihood.

Missing not at Random (MNAR)

The missingness of data is neither MAR nor MCAR, of which the value of the variable that is missing is related to the reason it is missing.

Naïve Bayes (NB)

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

Ontology-based method

The ontology-based method involves knowledge representation, data integration and management, semantic reasoning, information retrieval, and analysis. The ontology serves as the foundational framework for organizing information, data processing, or problem-solving for a specific domain.

Pre-demolition audit inventory

Pre-demolition audit inventory, also called waste audit inventory or environmental inventory, documents the presence and amount of hazardous substances that is used as a basis for construction and demolition waste management in renovation and demolition activities.

Polychlorinated biphenyls (PCB)

PCB is a mixture of chlorinated organic chemicals consisting of 209 congeners with no known taste or small and range in consistency from oil to waxy solid. Due to non-flammable, chemically stable, high boiling point, and electrical insulating properties, they were used extensively in industrial and commercial applications, such as plasticizers in building sealants.

Preferred reporting items for systematic reviews and meta-analyses (PRISMA)

PRISMA is an evidence-based checklist with 27 items used for improving transparent reporting in systematic reviews, which covers all aspects of the manuscript including title, abstract, introduction, methods, results, discussion, and funding.

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

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

Principal component analysis (PCA)

PCA 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.

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)

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

Pearson correlation

Pearson correlation indicates the degree of linear correlation between two variables yet does not imply causation. It gives information about the magnitude of the association, or correlation, and the direction of the relationship.

Receiver Operating Characteristic curve (ROC curve)

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier with varied discrimination thresholds where the true positive rate is plotted against the false positive rate.

Recall (REC, Sensitivity)

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

$$REC = TP / (TP + FN)$$

Spatial interpolation

Spatial interpolation is a method used in geostatistics and geographic information systems to estimate unknown values at certain locations based on known values at nearby locations. Such techniques are for example inverse distance weighting, kriging, splines, natural neighbor interpolation, trend surface analysis, and radial basis functions.

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.

Selective demolition

The removal of materials from a demolition site in a pre-defined sequence before demolition or renovation to maximize recovery and recycling performance.

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.

Soft Independent Modeling of Class Analogies (SIMCA)

SIMCA is a pattern recognition method that describes each class separately in a principal components (PC) space. Unknown samples are compared to the PCA class models and assigned to the class according to their analogy with the calibration samples.

1. Introduction

The urgency of transitioning toward a circular built environment to mitigate climate change and caution with resources cannot be overstated. In Europe, the construction sector is responsible for half of all extracted materials and total energy use, as well as one-third of water use and waste generation (European Commission, 2022). In the trajectory towards functional circular construction, the housing sector is implementing decarbonization and adaptation strategies, including renovation, reconstruction, and selective demolition, which are integrated with circular economy (CE) approaches. Such strategies aim to extend the service life of existing buildings and enhance their whole-lifecycle performance (Nußholz et al., 2023). However, the widespread presence of hazardous materials impedes progress in building material recovery and environmental impact reduction in the construction industry (Lewis, 2019; López Ruiz et al., 2020). Consequently, managing the risk of hazardous materials is crucial for achieving resource efficiency in circular construction practices (Bodar et al., 2018).

To manage the uncertainty associated with hazardous materials in situ, risk-based inspections are essential for the sustainable maintenance of both operational and end-of-life buildings (Kim et al., 2018). Pre-demolition audit is one of the examples, where contaminated components are identified to evaluate the quality of materials in construction and demolition waste (CDW) during collection and sorting processes (ECORYS, 2016). The environmental inventories created prior to reconstruction or demolition are vital to ensuring safe management of demolition waste by facilitating decontamination planning and waste handling schemes (Wahlström et al., 2019). Over the years, the accumulation of numerous environmental inventories has presented new opportunities to deepen our understanding of hazardous substance prevalence in existing building stocks. However, these data have not yet been digitally systemized nor explored in research.

1.1. Problem Statement

Unaligned Legislations on Reducing Material-derived Carbon

The increasing costs and scarcity of raw materials, coupled with objectives to reduce material-embodied carbon emissions, require immediate adoption of circular economy practices in the construction sector (European Commission, 2021). This shift is accompanied by updated regulations that emphasize efficient resource utilization and necessitate quality assurance for secondary materials in the construction market, which also impose new demands for hazardous materials management (Rašković et al., 2020). Key legislative developments are for instance, the Renovation Wave (European Commission, 2020b), the European Green Deal (European Commission, 2019), and the new Circular Economy Action Plan (European Commission, 2020a), all of which are expected to significantly impact a large portion of existing building stock. However, the major challenge identified is the lack of alignment between these EU policies and the EU REACH regulation (Registration, Evaluation, Authorisation and Restriction of Chemicals) (European Chemicals Agency, 2007). The disconnect lies in the separate governance of waste legislation and substance-specific legislation, which currently do not fully support emerging closed-loop initiatives (Bodar et al., 2018). The focus of REACH used to be the production and use phases, rather than the entire life cycle of chemicals, which inadequately addresses the prolonged use and eventual re-entry of hazardous components in buildings into waste streams. To facilitate safe and sustainable material reuse, frameworks for preliminary risk management have been proposed to assess potential hazards in reclaimed or recycled materials (Bodar et al., 2018). Nevertheless, further exploration is needed to predict, quantify, and characterize in situ hazardous substances in the building stock to support this transition.

Practical Challenges in the Construction Sector

To align with the circular economy trend, the construction industry faces several challenges. A significant hurdle is the limited understanding of the distribution and extent of hazardous materials remaining, which impedes effective hazardous material management from passive mitigation measures to proactive decontamination (Bergmans et al., 2017; ECORYS, 2016; Rašković et al., 2020; Wahlström et al., 2020). The unforeseen discovery of hazardous substances in decommissioned construction products introduces uncertainties in cost and schedule planning (Wahlström et al., 2019), and poses risks to occupational health (Cook et al., 2022) during renovation and demolition activities. Although most hazardous substances are now banned or strictly regulated, the risk of exposure persists globally (Cook et al., 2022). Therefore, developing new tools or guidelines for risk mitigation and identification of hazardous materials in CDW streams, both during and post-construction or reconstruction, is imperative (Arevalillo et al., 2017). Understanding the approximate presence of in situ hazardous materials can enhance

material quality management and traceability systems. This knowledge is crucial for facility management in predictive maintenance and in the construction industry, particularly in semi-selective demolition processes, to prevent secondary contamination of CDW (Bergmans et al., 2017; Wahlström et al., 2019).

1.2. Theoretical Framework

This thesis is grounded in the following knowledge domains: construction and demolition waste management, urban informatics, and the application of artificial intelligence, as depicted in Figure 1.1. The primary objective is to explore the feasibility of developing data-driven methodologies for predicting the presence of hazardous materials in situ within the construction sector. This involves leveraging the potential of applied AI to enhance urban informatics strategies, specifically in the context of construction and demolition waste management.

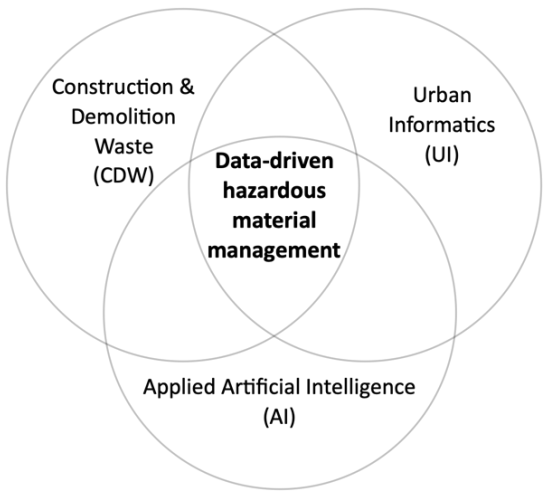


Figure 1.1. Interdisciplinary research scope.

1.2.1. Research Scope

Applied AI is used as as a key leveraging technique for hazardous material assessment, utilizing extensive data derived from urban informatics. The knowledge domains underpinning the study are elaborated and exemplified below.

Construction and Demolition Waste (CDW)

Efficient management of Construction and Demolition Waste (CDW) is crucial for enhancing material recovery rates and closing resource loops with huge potential for carbon footprint reduction according to the EU Circular Economy Action Plan and the revised Waste Framework Directive (2008/98/EC, amended 2018/815) (Wahlström et al., 2020). With nearly 40% of total waste generated by the construction sector (European Commission, 2022), the need for effective CDW management strategies concerning waste prevention, reuse, and recycling is underscored. Key barriers identified included: ineffective CDW regulation and insufficient economic incentives from policy measures, underdeveloped reverse logistics and dissembly or remanufacture infrastructure, quality concerns of recovered components resulting in low readiness of secondary material market, lack of comprehensive business models and marketplace, insufficient information on historical product usage in buildings (Bergmans et al., 2017; Sandberg & Hultegård, 2021; Villoria Sáez & Osmani, 2019; Wahlström et al., 2020).

Meanwhile, practical CDW management protocols have been developed at both EU and national levels to refine pre-demolition audits and waste handling processes (ECORYS, 2016; Wahlström et al., 2019). In Sweden, a tailored resource and waste guideline was established by the construction industry to inform pre-demolition audit inventories (Byggföretagen, 2019). Suggested future research avenues include pollutant identification and control in CDW for enhanced recyclability, performance evaluation and lifecycle traceability of CDW products, and the efficient integration of information technology in CDW management (Ajayi et al., 2015; Wu et al., 2019).

Urban Informatics (UI)

Urban informatics (UI) is an emerging field that focuses on understanding, managing, and designing systems in built environment through computational approaches (Shi et al., 2021). This field has gained momentum with the advancement of digitalization, leading to a paradigm shift in traditional building stock research towards more quantitative analyses (Kohler, 2018; Shi et al., 2021). Urban informatics encapsulates the dynamic interplay among various urban components, such as morphology, mobility, space-time patterns, energy and infrastructure systems, and spatial economics, by integrating systems theories and methods from urban science, geomatics, and informatics (Shi et al., 2021). Nowadays, 55% of the global population resides in urban areas and the figure is projected to rise to 68% by 2050 (UN Habitat, 2022). This rapid urbanization

underlines the critical need for addressing anthropogenic greenhouse gas (GHG) emissions by redeveloping carbon-neutral urban infrastructure and building stocks (Chen et al., 2023). In response to new legislation and resource constraints, contemporary research in building stock is increasingly focused on the decarbonization of built environments. This includes improving energy efficiency and promoting material reuse and recovery in construction (Chen et al., 2023).

Urban informatics fosters the creation of synergistic solutions for complex urban challenges by integrating stakeholders' needs with multidimensional data from various sectors (Shi et al., 2021). This approach is operationalized through multi-scale methodologies: building information modeling (BIM) and lifecycle analysis (LCA) at the building scale; material flows and stock analysis (MFA) at the regional level; and remote sensing and geographic information systems (GIS) at the continental scale. The application of these modeling techniques, in conjunction with data from digital infrastructures, enables comprehensive management of building stocks, addressing social, environmental, and economic aspects (Carbonari et al., 2019; Koutamanis et al., 2018; Lucchi et al., 2018).

Applied Artificial Intelligence (AI)

Applied artificial intelligence (AI) refers to the utilization of intelligent, autonomous, and purpose-driven computational systems for data-centric problem-solving and decision-making (Darko et al., 2020). Under the umbrella term of AI, key subfields can be categorized into expert systems, agent systems, and machine learning (ML) (Norvig & Russell, 2021). Expert systems utilize if-then rules to facilitate decision-making, avoiding the use of procedural codes (Węglarz & Gilewski, 2017). When presented with new queries, these systems leverage an inference engine to extract applicable rules and facts from knowledge bases, thereby deducing new information. In contrast, agent systems deploy multiple intelligent entities that autonomously interact with their environment using sensors and actuators to accomplish specific objectives (Xiang et al., 2022). Machine learning differs from these systems by relying on mathematical and architectural models, representing knowledge tasks related to pattern recognition (Raschka & Mirjalili, 2019). Common AI techniques widely applied in the architecture, engineering, and construction (AEC) industry including genetic algorithms for structural and design optimization, neural networks for predicting structure strength, convolutional neural networks for damage detection, fuzzy logic for uncertainty assessment, and machine learning for system identification and structural health monitoring.

Machine learning, in particular, has gained traction in building research due to its capability to autonomously learn from data, progressively improving by learning from errors without explicit programming. Common ML functionalities include supervised classification and regression (using decision-tree, support vector machine, naïve Bayes, and neural network methods), unsupervised clustering and dimensional reduction (employing K-means, DBSCAN, and hierarchical clustering methods) (Yan et al., 2020). Literature has shown promising potential for ML to

enhance building performance throughout its lifecycle (Hong et al., 2020). Predominant applications of ML in building operation and maintenance involve fault detection, diagnostics, energy efficiency, post-occupancy evaluation, and building control for energy savings, grid interactivity, and comfort enhancement (Hong et al., 2020). Additionally, ML is increasingly being deployed in building retrofitting for identifying retrofit potentials, evaluating energy-saving measures, characterizing buildings, and in building design for parametric design and evaluation. Its application extends to construction, encompassing cost optimization analysis, construction management, defect detection, Building Information Modeling (BIM), and CDW management.

1.2.2. Interdisciplinary Field

Data-driven hazardous material management

Data-driven hazardous material management focuses on assessing the risk of contamination in existing buildings by leveraging building-specific environmental data and comprehensive building databases. Since the 1970s, global bans on asbestos-containing materials (ACM) and PCB (Polychlorinated Biphenyls)-containing products have been enacted, accompanied by national strategies for intervention in most countries (Westerholm et al., 2017, Kim and Yu, 2014). Similarly, the recognition of indoor radon as a health hazard has led to regulations on radioactive materials (Copes & Peterson, 2014; Kim & Yu, 2014). Despite the prohibition of these hazardous materials for decades, their widespread historical use in construction poses ongoing occupational and public health risks. For instance, asbestos is responsible for approximately 250,000 deaths annually worldwide (Westerholm et al., 2017), PCB exposure has been linked to increased mortality (Parada et al., 2020; Ruder et al., 2014), and indoor radon is estimated to cause 3-14% of lung cancers globally each year (Cook et al., 2022).

Moreover, the growing trend of reusing materials from CDW in circular construction raises stringent requirements for contaminant-free materials. Prioritizing risk mitigation, control, and monitoring of hazardous exposure, research focuses on the safe management and optimal disposal of these materials. Previous studies have concentrated on identifying the sources, pathways, types, and quantities of asbestos (Kim & Hong, 2017; Nam et al., 2015; Song et al., 2016; Franzblau et al., 2020; Govorko et al., 2019; Mecharnia et al., 2019) and PCB-containing materials (Diefenbacher, 2016; Shanahan et al., 2015; Herrick et al., 2016; Diamond et al, 2010) through field and laboratory investigations. On the other hand, indoor radon research has characterized influencing factors and their association with active monitoring (Cerqueiro-Pequeño et al., 2021; Valcarce et al., 2022), as well as the forecast of long-term indoor radon concentration trends (Kropat et al., 2015a; Kropat et al., 2015b; Elío et al., 2019). Currently, challenges in data-driven

applications for CDW include poor data quality and limited representation of knowledge in case studies (Yan et al., 2020). Future research is suggested to develop a holistic data mining framework, extract knowledge from unstructured data, and evaluate advanced data mining techniques to address these challenges (Yan et al., 2020).

1.3. Previous Research

The identification and quantification of in situ hazardous substances are crucial for effective building stock decontamination and achieving material circularity. Current CDW management highlights the importance of traceability of hazardous materials to mitigate potential disturbances and emissions during building operation and retrofitting phases (Wu et al., 2019). Despite established EU and national guidelines for CDW management, adequate and precise information on the building structure and building material constitution for exiting buildings is missing (Rašković et al., 2020). As a consequence, various sources of demolition-related information are employed for decision support in renovation and deconstruction planning, including building permit documentation, drawings, field survey records, BIM models, etc. (Rašković et al., 2020).

Machine learning and statistical techniques are utilized to impute missing values and generate insights from incomplete, finite information. For instance, Yang et al., (2021) developed aggregated behavior-based ML models from waste generation behaviors to handle project-level missing not at random data, yielding satisfactory results. Mecharnia et al. (2019) predicted the presence of asbestos materials using an ontology-based approach with incomplete temporal data on marketed products. So far, the potential of using waste audit inventories as input data for training ML models remains unexplored and unsystemized. The reliability of waste audit inventory heavily depends on available documentation and field survey results, making the distinction and volume estimation of hazardous materials uncertain (Rašković et al., 2020). Therefore, it is beneficial to compile environment-specific data, integrate them into large-scale building databases, and determine their utility for contamination assessment.

In addition to data extraction, digital tools and methodologies for estimating waste stream types and volumes at the urban scale are proposed to support sustainable CDW management and the selection of appropriate urban mining strategies (Powell et al., 2015; Rašković et al., 2020). A primary focus of ML-related studies in the CDW domain is on predicting waste generation (Akanbi et al., 2020; Cha et al., 2017, 2020; Yu et al., 2019) and the reuse of building components (Deepika et al., 2022). However, these ML applications do not facilitate quality evaluation of in situ building materials, leaving the characteristics of remaining hazardous materials unknown. Conducting comprehensive pre-demolition audits is

resource-intensive, as many hazardous substances are not visually distinct and require individual building-based in situ sampling and laboratory analysis (Powell et al., 2015). Furthermore, since inventory results inform project management planning and tendering of disposal services, deviations from actual waste types and quantities can incur financial and environmental costs for contractors and property owners (Franzblau et al., 2020; Rašković et al., 2020). Thus, it is of great necessity to ensure accurate information on the waste audit operations meanwhile expanding the capacity of hazardous material screening (Rašković et al., 2020).

In Appendix I, literature related to data-driven applications for building contamination prediction is summarized. A comprehensive inventory of urban asbestos and PCB is a critical first step in identifying and quantifying their occurrence in the built environment for contaminant removal or remediation (Glüge et al., 2017). Various data sources have been explored for this purpose and analyzed using statistical or machine learning techniques. These include registers of buildings undergoing asbestos abatement from demolition databases (Franzblau et al., 2020), ACM characterization from landfills (Powell et al., 2015), self-reporting of recognized ACM via a mobile app (Govorko et al., 2017, 2018, 2019), field inventory and surveys on historical plants of ACM (Wilk et al., 2015, 2017, 2019), remote sensing data on ACM (Krówczyńska et al., 2020; Raczko et al., 2022), air sampling for PCB (Diamond et al., 2010; Diefenbacher et al., 2016; Kolarik et al., 2016; Robson et al., 2010a), and indoor radon measurements (Adelikhah et al., 2021; Elío et al., 2019; Khan et al., 2021; Kropat et al., 2015a; Kropat et al., 2015b; Oni et al., 2022; Sarra et al., 2016; Valcarce et al., 2022). Few studies adopt ML approaches to predict ACM and indoor radon, with most research remaining descriptive in hazardous substance characterization from a bottom-up perspective. On the contrary, material flow and material stock analyses offer a top-down perspective on hazardous material metabolism, enabling impact assessment of hazardous substances.

However, both top-down and bottom-up approaches exhibit limitations in accurately estimating hazardous material in building stocks at an urban scale. Previous analyses of stocks and flows have been plagued by significant uncertainties, particularly regarding the detailed composition of components and their lifespans (Bergsdal et al., 2014; Donovan & Pickin, 2016a). Aggregated data, encompassing annual production, import, and historical usage of hazardous materials, relies on various assumptions, leading to considerable variability in sensitivity analyses (Bergsdal et al., 2014; Donovan & Pickin, 2016a). Conversely, field sampling methods, while detailed, demand substantial resources for widespread implementation and are generally limited to airborne contaminants (Shanahan et al., 2015). Hence, the results from atmospheric chemical transport models may not comprehensively represent the entire building stock. Additionally, the geographic variability of certain hazardous substances warrants attention, as the level of exposure significantly fluctuates based on the age and type of buildings in question (Robson et al., 2010b). Given these challenges, data-driven approaches emerge as

promising alternatives, offering the potential to identify patterns of hazardous substances through historical inspection records and to estimate their likelihood of presence in buildings that have not been inspected.

1.4. Research Gaps

The absence of scalable, cost-efficient methods to systematically determine the probability of in situ hazardous substances in existing buildings impedes risk assessment and the planning of safe and cost-effective renovation and demolition.

Gap 1 Assessing the information availability and usability of pre-demolition audit inventories.

As part of the EU Construction and Demolition Waste Management Protocol, pre-demolition audits have become either mandatory or a partially voluntary practice for renovation and demolition permit applications in many European countries. Yet, the environmental information amassed over the years remains largely untapped due to a lack of standardization and digitalization, resulting in challenges in combining it with from other building registers. Therefore, harnessing the potential of these inventory data to characterize and estimate the residual hazardous substances in the building stock presents a significant opportunity.

Gap 2 Investigating data-driven techniques for predicting contamination in the built environment.

To date, machine learning has been underutilized in identifying hazardous substances within the building stock. Previous research in this area has predominantly utilized statistical methods for descriptive purposes rather than predictive analytics. Consequently, the specific challenges and opportunities of using data-driven approaches in this context remain largely unexplored.

Gap 3 Enhancing knowledge and awareness of in situ hazardous substances for circular building stock management.

Current knowledge about the presence of in situ contaminants is incomplete and difficult to confirm, complicating environmental inspections, selective demolition, and waste sorting. Concerns about hazard exposure add uncertainty to project timelines and cost estimates, as well as the safe management of demolition waste and remediation efforts. To advance circular construction practices and improve the quality of CDW, there is a need to develop extensive knowledge about the patterns of hazardous material presence and the extent of contamination within the building stock.

1.5. Research Focus

Aim The primary goal is to explore the potential of applying data analytics and machine learning in predicting and interpreting the presence patterns of in situ hazardous substances in existing building stock as a decision support for relevant actors.

The research questions are sequentially interrelated as fundamental elements of the machine learning pipeline, aligning with the previously identified research gaps. RQ1 focuses on assessing the availability and quality of environmental inventory data and identifying suitable target materials for machine learning predictions. RQ2 delves into the identification of appropriate modeling techniques considering data constraints and evaluates model performances. RQ3 is a rather open question exploring how the developed models and discerned patterns can facilitate the estimation of residual hazardous substances for risk-based inspection planning prior to renovation and (selective) demolition.

RQ1 Data gathering

What is the potential of using data from building registers and pre-demolition audit inventories for mapping hazardous substances in the building stock?

RQ2 Method development

Which predictive methods can be used to estimate the presence of hazardous substances in buildings?

RQ3 Prediction results

With what certainty can the presence of hazardous substances in the building stock be predicted?

Figure 1.2 below illustrates the thesis structure for two thematic tracks, employing three data-driven methodologies. Appended papers are grouped according to the research questions they address. Research works focusing on hazardous materials such as asbestos and PCB are indicated in blue, while those concentrating on radioactive substances such as radioactive concrete and indoor radon are highlighted in brown. This consistent color scheme is applied across the thesis for ease of navigation and comprehension. Paper I corresponds to RQ1 and offers an extensive overview of data-driven applications in managing in situ hazardous materials. It reviews state-of-the-art literature to establish a theoretical framework for the study. Paper II and its preceding study, Conf I, elaborate on RQ1, detailing the data gathering process for acquiring input data from environmental inventories of CDW and building registers. These empirical works aim to create a digital hazardous material dataset from pre-demolition audit inventories. Paper III-VI and Conf II concern RQ2 and RQ3, with specific emphasis on hazardous

materials and radioactive substances, respectively. Paper III serves as a pilot study to test ML applications in predicting asbestos and PCB-containing materials. Paper IV extends the work of Paper III, broadening the prediction scope to various hazardous materials across multiple building types. Paper V is an independent study examining a statistical learning method to estimate the presence of radioactive concrete (so called blue concrete) in major Swedish cities, based on inventory records of demolished and renovated buildings. Paper VI and Conf III address the prediction of indoor radon using varied ML and statistical modeling approaches.

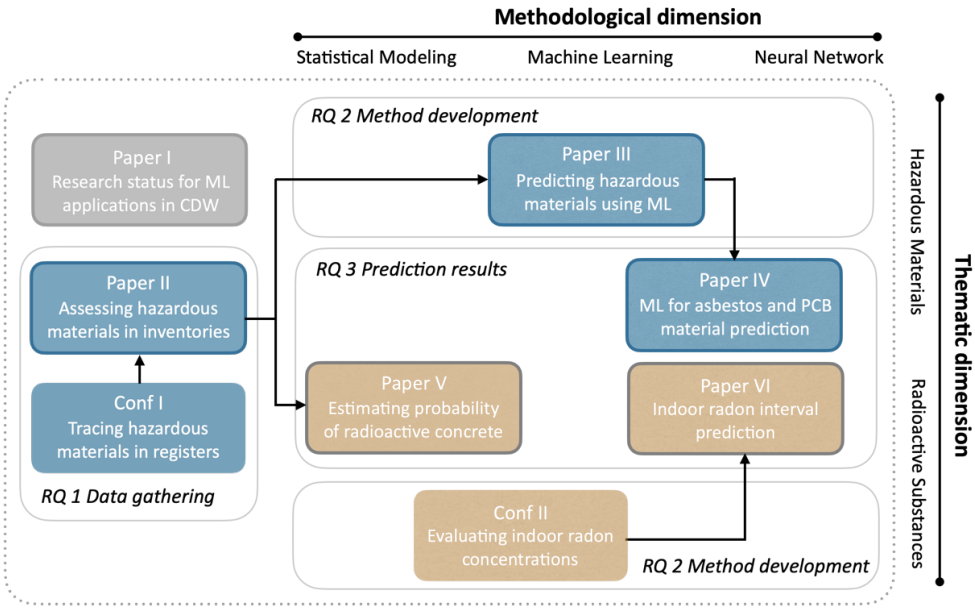


Figure 1.2. Thesis structure with thematic and methodological dimensions.

1.6. Liminations

The thesis has three primary delimitations concerning data, specifically focusing on (i) the quality of inventory data, (ii) the availability of key data features, and (iii) assumptions regarding building typology. Each delimitation is accompanied by a thorough description of its causes and the data processing techniques employed to mitigate their impacts:

Data quality of inventories

The thesis involves textual knowledge discovery by transforming unstructured empirical data from the CDW field into systemized digital data. The data quality of pre-demolition audit inventories was evaluated in terms of accuracy, completeness, reliability, relevance, and timeliness. Significant efforts were made to digitize these inventories accurately, converting inspection records into a coherent hazardous material dataset. A cross-validation workshop was conducted with the project team to ensure consistent documentation and agreement on the interpretation of detection records. Despite efforts to maintain completeness, some missing values are unavoidable due to the scope of inspections and inventory details. Lacking descriptions of generic characteristics were supplemented with information from additional sources, such as building registers and property maps, though the state of uninspected building components remained unknown.

The reliability of data was gauged considering the inspectors' experience and competence levels and the number of samples collected, yet verification of inspection records was not feasible, posing a risk of judgment errors. The granularity of data varied, depending on whether the source of inventory was substance-level (e.g., demolition and control plans) or material-level (e.g., protocol templates and reports). The dataset design only included relevant and frequently available information to ensure appropriate and sufficient hazardous substance documentation. The constantly changing state of the building stock and the potential mismatch between inspection records and building register data posed challenges in assessing the timeliness of inventories. With the alteration of existing building stock, old building registers may be eliminated or renewed, leading to increasing uncertainty as opposed to the date of inventories.

Data availability of key features

The study aims to develop a comprehensive hazardous substance prediction pipeline, utilizing general building characteristics as training features. Nevertheless, the occurrence patterns of certain hazardous substances are specific to component use and building type. For example, PCB capacitors and sealants are prevalent in large non-residential buildings with high electricity demand (Diamond et al., 2010; Shanahan et al., 2015), and the likelihood of detecting ACM increases with building age and building physical footprint (area per floor) (Song et al., 2016). To refine

model training, it was necessary to include relevant features tailored to specific hazardous substances. Other potential critical features, such as construction materials and proximity to hazardous material production plants, were suggested in the literature (Kropat et al., 2014; Wilk et al., 2015) but were not always available in building databases, limiting the predictive performance of ML models. Information such as foundation types, crucial for indoor radon concentration predictions, was only accessible in selective municipal databases. As such, features not universally available were excluded from the national-scale analysis.

Assumptions on building typology

Aggregating buildings with similar typologies is a way to simplify the complexity in the building stock analysis, assuming that buildings with similar functions and ages in neighboring areas have particular constitutions (Berggren & Wall, 2019). The building stock was clustered into different types based on location, construction year, and usage, with the dataset's partition level depending on the data size of dependent and independent variables for prediction. To enable an accurate inference of the analytical results, observations were divided into subgroups for modeling based on building category and usage codes from municipality cadastral registers. The primary categorization of building types is into residential and non-residential dwellings, further subdivided into ten distinct building classes for detailed analysis. Machine learning models were developed for each building class subgroup, with other categorical variables such as municipality or postcode employed as dummy features for model training and interpretation.

1.7. Content Structure

The thesis is structured into three main sections - the prologue, main body, and epilogue. The prologue offers an overview of the research topic, followed by a detailed introduction to the data and digital techniques employed in the study. This is encompassed in Chapter 1 Introduction and Chapter 2 Materials and Methods. The main body, represented by Chapter 3, presents key research findings, drawn from appended papers and a comprehensive report, segmented into four distinct sections. The thesis concludes with the epilogue, comprising Chapter 4 Discussion and Chapter 5 Conclusions, which collectively encapsulate the study's final insights and findings.

2. Materials and Methods

The chapter offers a comprehensive overview of the materials and methods used in the thesis. Employing the framework of Knowledge Discovery in Database (KDD), the materials are depicted in gray and the methods in white, segmented into three successive parts in Figure 2.1. Section 2.1 elaborates on qualitative approaches, aligning theoretical and practical knowledge of the research domain through a systematic literature review and the collection of industry insights. Section 2.2 focuses on the empirical aspect, detailing the process of data gathering and analysis of pre-demolition audit inventories from buildings that underwent renovation or demolition between 2010 and 2022. It describes the procedures for assembling and validating the database, evaluating it, and conducting exploratory data analysis. Section 2.3 presents quantitative approaches for predictive modeling, patterns evaluation, and predictive analysis. The selection of relevant data analytic and modeling techniques is tailored to the objectives of the analysis, considering data quality, quantity, and the desired predictive performance.

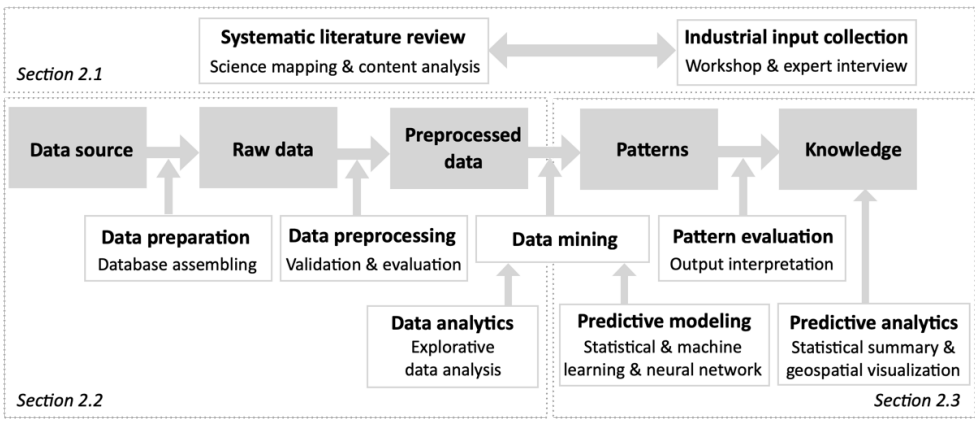


Figure 2.1. Materials and methods illustrated based on the framework of Knowledge Discovery in Database.

Various elements of the KDD framework were related to the material and methodology used in the papers, as depicted in Figure 2.2. Paper I reviewed academic publications and gray literature using science mapping and content analysis to explore state-of-the-art research developments. Paper II and Conf I

merged building registers and environmental inventories, assessing their data quality using data analytic and statistical methods. Papers III and IV further performed descriptive analysis (such as summarization, clustering, and association) and predictive analysis (utilizing machine learning classification models) on the hazardous material dataset. This was done to discern patterns, identify correlations, and evaluate the likelihood of asbestos and PCB presence in buildings. Paper V deduced the occurrence of radioactive concrete by analyzing detection records from environmental inventories and indoor radon measurements using a learning Bayesian network. Finally, Paper VI and Conf II resembled the data-driven modeling approach and pipeline to predict indoor radon concentrations and intervals in the Swedish building stock.

	Material	Method
Paper I Systematic literature review	Article/conference proceeding /review paper	Science mapping & content analysis
Paper II / Conf I Data assembling and evaluation	Building register & environmental inventory	Data preparation & data preprocessing & data analytic
Paper III-IV Hazardous material prediction	Building register & environmental inventory	Data analytic & predictive modeling & pattern evaluation & predictive analytic
Paper V Radioactive concrete estimation	Building register & environmental inventory & indoor radon measurement	Data analytic & predictive modeling & pattern evaluation & predictive analytic
Paper VI / Conf II Indoor radon prediction	Building registers & indoor radon measurement	Data analytic & predictive modeling & pattern evaluation & predictive analytic

Figure 2.2. Representation of materials and methods in the appended papers.

2.1. Qualitative Approaches

Figure 2.3 illustrates the alignment of knowledge in hazardous material management, bridging the gap between theoretical frameworks and practical applications. This was achieved by conducting a systematic literature review, which outlined the theoretical evolution in the field and identified existing gaps and limitations in current studies. Concurrently, inputs from relevant industry stakeholders were gathered to verify their research needs and priorities. Aligning theoretical insights with practical realities is crucial for steering the research in a direction that enhances the practical applicability and utility of the research findings.

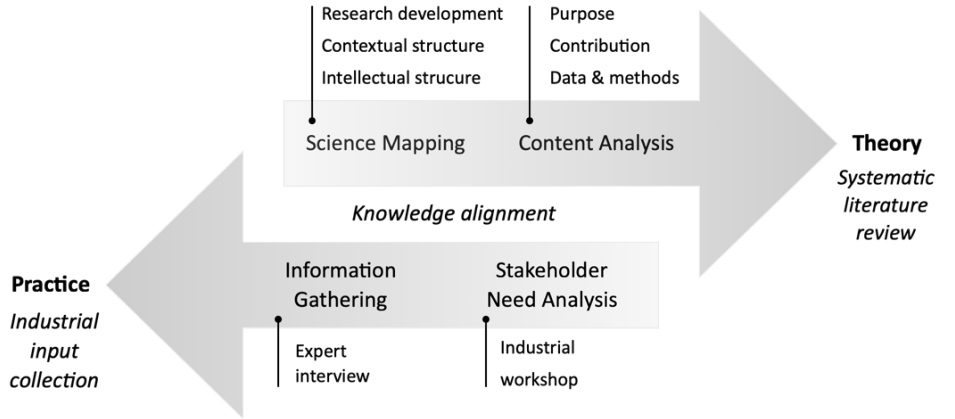


Figure 2.3. Outline of qualitative study approaches.

2.1.1. Systematic Literature Review

In Paper I, a two-fold approach was employed for the systematic literature review. The first part involved assessing the scientific progress in the knowledge domain, while the second part focused on identifying data-driven approaches and their input data in relevant studies. The literature search was conducted using Web of Science and Google Scholar, utilizing a combination of search phrases such as “hazard”, “(artificial intelligence) AI or machine learning”, and “building” linked by Boolean operators. This process resulted in the retrieval of 307 documents, which were analyzed using the R library Biblioshiny for bibliometric analysis (University of Naples Federico II, 2023). The meta-data of the acquired literature was examined to describe the research development, conceptual structure, and intellectual structure of the field in the first phase of science mapping (Chen, 2017). Research development was quantified by the number and thematic distribution of scientific publications. Subsequently, the conceptual structure was clarified through co-word and word dynamic analysis, using multiple correspondence analysis (MCA) (STHDA, 2017) to illustrate the relationships and evolution among keywords. The historiographic mapping was depicted in citation networks and three-field plots (Garfield, 2004), showcasing the knowledge contributions from authors and their publication outlets.

The second part of the paper, the content analysis, utilized the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Moher et al., 2009) to gauge the relevance of the literature for more detailed analysis. Through this funnel-like process, 57 papers with middle to high relevance regarding machine learning applications in CDW management, asbestos, and PCB topics were selected for critical literature review. From these, 16 highly relevant papers were synthesized following the activity flows in the EU Construction and

Demolition Waste Management Protocol (ECORYS, 2016). This synthesis summarized the state-of-the-art machine learning applications and input data used for identifying hazardous materials, source separation, and on-site collection.

2.1.2. Industrial Input Collection

To gather industrial insights from the construction industry on our research development, a dialogue was initiated with actors from the Swedish CDW sector through a workshop with stakeholders and expert interviews. The workshop was designed to facilitate collaborative discussions among construction industry participants to obtain valuable feedback (Lain, 2017), whereas expert interviews were conducted to deepen the knowledge of pre-demolition audits and decontamination practices. Expert interviews focused on delving deeper into pre-demolition audits and decontamination practices. This approach bridged the gap between theoretical research and practical application by combining diverse stakeholder perspectives with specialized expert knowledge. A workshop titled "Identification of hazardous materials with applied AI" was conducted in November 2022 with two discussion sessions and an intermediary presentation. The structure of the workshop was as follows:

- Introduction to the research topic (10 min)
- Part I: Research question answering and open discussion (40 min)
- Presentation of preliminary research results (15 min)
- Part II: Gathering feedback on the research (10 min)
- Discussion of next steps (10 min)

Table 2.1 below details the sectors, positions, and expertise of the 13 workshop participants. A diverse and balanced representation from various sectors of the CDW value chain was a key consideration in participant selection. The outcomes of the workshop were then used for a stakeholder needs analysis, which helped to understand the problem's scope and identify key actors interested in utilizing the predictive models. This same network was later leveraged for disseminating the research findings in December 2023, as described in section 4.3.2.

Table 2.1. Overview of the participants in the workshop.

Sector	Position	Field of expertise
Academic	Senior researcher	Sustainable renovation
Academic	Researcher	Building stock research
Academic	PhD candidate	Circular construction, applied AI
Construction firm A	R&D manager	BIM, digitalization
Consultancy firm B	Building specialist	Material inventory, applied AI
Consultancy firm B	Technical manager	Inventory, circular construction
Consultancy firm B	Procurement	Demolition, recycling
Consultancy firm C	Senior consultant	Pre-demolition audits
Consultancy firm C	Senior consultant	Pre-demolition audits
Consultancy firm D	Senior consultant	Pre-demolition audits
Decontamination firm	Consultant	Circular construction
Client	Senior project leader	Demolition development
Housing authority	Senior project leader	Circular construction, digitalization

Thereafter, a semi-structured interview was scheduled with an in-house expert to gather detailed insights on pre-demolition audits and decontamination practices. The problem-centered interview was regarded as an effective method for delving into implicit and in-depth knowledge within a specific area (Döringer, 2021). It employed an interactive-dialogue interview approach guided by a pre-defined interview protocol, aiming a deep exploration of the interviewee's individual perspectives. In January 2023, an expert interview was conducted with a senior environmental engineer, who has over 16 years of combined industrial and research experience. Prior to the interview, a semi-structured questionnaire, divided into two sections encompassing 11 questions, was shared with the expert. The first part of the questionnaire addressed pre-demolition audits, while the second focused on decontamination processes. This was followed by a comprehensive one-and-a-half-hour online discussion, which allowed for an in-depth exploration of the topics. The findings from both the stakeholder need analysis and this expert interview were systematically compiled and are presented in Section 3.1.2.

2.2. Data Gathering and Analytics

Data curation is an iterative process that involves selection and organization of potential features into databases for specific predictive analyses. For this study, various environmental information sources concerning asbestos, PCB, radioactive concrete, and indoor radon were screened, gathered, and analyzed to trace the presence of hazardous substances in the Swedish building stock. As depicted in Figure 2.4, two distinct databases were created: one for hazardous materials and another for indoor radon. These databases were assembled by integrating building-

specific environmental data with generic building stock data, thereby establishing a robust foundation for model training and predictive tasks. A key step in this process was the use of the national real estate index, combined with building addresses, as primary matching keys. This approach enabled the creation of comprehensive datasets with high granularity, pinpointing the location of identified hazardous substances at the individual building level. The methodologies and specifics of this data gathering workflow are detailed in Paper II and Conf I for the hazardous material database, and in Paper VI for the indoor radon database. These descriptions provide a clear insight into the procedures for database compilation and their subsequent application in hazardous materials and substances prediction.

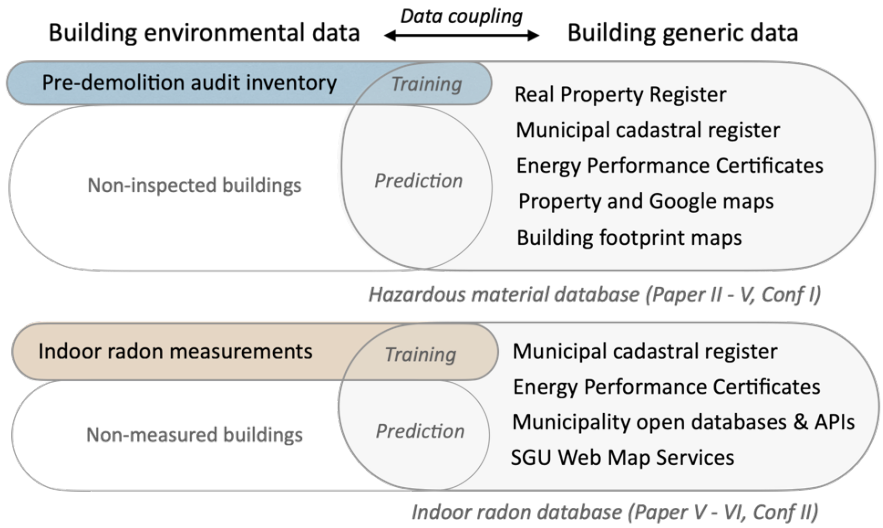


Figure 2.4. Data coupling between building specific and generic data.

2.2.1. Database Curation

The hazardous material database and the indoor radon database were created using Feature Manipulation Engine (FME) and Python for statistical and machine learning modeling to estimate remaining hazardous substances in the building stock that lack existing inventories or indoor radon measurements using a method established by Johansson et al. (2017).

Hazardous Material Database

In compliance with the EU Waste Framework Directives (European Commission, 2008), carrying out a pre-demolition audit is a mandatory or conditionally voluntary practice in EU countries. These audits include a desk study of building documentation and maintenance protocols, field surveys for materials and

substances identification, classification, and sampling, along with management recommendations and reporting. In Sweden, these environmental inventories are essential for building permit applications for renovation, reconstruction, extension, and demolition (Grönberg, 2017). However, they are not included or integrated in the national building databases due to a lack of systemization and digitalization.

The created hazardous material database collates building environmental data from these pre-demolition audit inventories with generic building data. The database contains approximately 1,100 observations of environmental and PCB inventories from buildings renovated or demolished between 2010-2022 in Stockholm, Gothenburg, Malmö, and Kiruna municipalities. The initial case study in Gothenburg in 2020, presented in Conf I and Paper II, was followed by expanded searches in Stockholm (Paper III) in 2021, and later in Malmö and Kiruna (Paper IV) between 2022-2023.

To standardize and compile various hard-copy inventory records into a consistent template, a protocol template was created based on the format introduced by the City of Gothenburg. This protocol included information on inspection extent, completeness, and detected hazardous materials. The documentation in reports was detailed based on the list of hazardous waste from the resource and waste guidelines and thorough with lab analysis of material samples, whereas control and demolition plans typically had simpler formats with free-text descriptions of detected substances. Consequently, a list of common hazardous substances and materials was included in the dataset, covering various components such as mercury, CFC/HCFC, PCB, asbestos, PVC, and radioactive concrete. Unlisted materials were categorized under “other components”. This approach balanced the need for sufficient observational data without losing crucial details. A cautious principle was followed in the digital transformation to ensure correct documentation of affirmed detection versus unknown or unsure records of hazardous materials through an internal workshop of transforming ten observations with the research team.

The building database comprises data from the Swedish real estate taxation register, municipality cadastral register, Energy Performance Certificates (EPCs), and building footprint maps from the metropolitan regions of Stockholm, Gothenburg, and Malmö (Swedish Land Survey, 2022), detailed in Table 2.2. The matched registers of inventoried buildings were appended to the inventory template for model training, while the remaining building registers were utilized for model predictions. Google Maps and property maps served as supplementary data sources for building validation during data preprocessing and merging. The step-by-step process of creating the hazardous material dataset is outlined in Conf I Table 1, with detailed data specifications provided in Paper II Table 1. Each observation in the database includes comprehensive information on building cadastres and characteristics, types and quantities of hazardous components, and building materials. Machine learning models for asbestos- and PCB-containing materials, developed from this database, are described in Papers III-IV. Additionally, the radioactive concrete dataset was partly derived from the hazardous material

database and partly from the indoor radon measurements in the municipalities of Gävle and Umeå. Detailed data specification of the radioactive concrete dataset and learning Bayesian network modeling are described in Paper V.

Table 2.2. Overview of attributes in the hazardous material database.

Source	Aggregation	Attribute*	Variable type
Pre-demolition audit inventory	Building	National property index	Matching key
		Address	Matching key
		Building type[1]	Independent
		Building usage	Ancillary
		Construction year[1]	Independent
		Renovation year[1]	Independent
		Renovation extent	Ancillary
		Floor/inventory area[1]	Independent
		Inventory types	Ancillary
		Scope	Ancillary
		Inventory date	Ancillary
		Auditors	Ancillary
		Decontamination	Ancillary
		Asbestos detection	Dependent
		Asbestos components	Dependent
		PCB detection	Dependent
		PCB components	Dependent
		Radioactive concrete	Dependent
		Radioactive concrete component	Dependent
Municipal cadastral register	Building	National property index	Matching key
		Postcode	Independent
		Postal location	Ancillary
		Building category code	Independent
		Building type code	Independent
		Construction year[4]	Independent
		Floor area[4]	Independent
Swedish Real Estate Taxation Register	Property	National property index	Matching key
		Property type[2]	Independent
		Construction year[2]	Independent
		Renovation year[1]	Independent
		Floor area[2]	Independent
Energy Performance Certificates	Building	National property index	Matching key
		Address	Matching key
		EPC approved date	Ancillary
		EPC building category	Independent
		EPC building type	Independent
		Construction year[3]	Independent
		Renovation year[2]	Independent

Source	Aggregation	Attribute*	Variable type
Energy Performance Certificates	Building	Heated floor area[3]	Independent
		Number of floors	Independent
		Number of apartments	Independent
		Number of stairwells	Independent
		Number of basements	Independent
		Ventilation types	Independent
Building footprint map	Building	Building physical footprint[5]	Independent
		Number of floors	Independent

*Variables marked with superscripts were found in multiple registers. The numbers signified quality ranking when consolidating multiple entries into a singular record.

Indoor Radon Database

In 1981, Sweden's first indoor radon reference limit of 200 Bq/m^3 was introduced as part of the national radon program (Rönnqvist, 2021). Subsequently, several extensive indoor radon surveys have been conducted, including the ELIB survey between 1991-1992, the Radon Survey in the 2000s, and the BETSI survey from 2007-2009. These surveys have focused on residential buildings, yielding a significant number of long-term average indoor radon measurements: 340,000 in single-family houses and 440,000 in multifamily houses (Rönnqvist, 2021).

The indoor radon database encompasses a broader range of attributes and a larger dataset compared to the hazardous material database, as detailed in Table 2.3. It includes anthropological and geographical data from the Swedish Cadastral and Land Registration Authority (Swedish Land Survey et al., 2022), i.e., building parameters and coordinates, as well as geological data from the Geological Survey of Sweden's (SGU) Web Map Service databases (Swedish Land Survey et al., 2022), i.e., radioactive substance concentrations and soil types, as training features. By merging these data with the nationwide indoor radon measurements from the latest EPCs and municipality open databases, the indoor radon database was created with around 190,000 measured properties between the 1990s and 2021.

Afterward, the annual average indoor radon levels, derived from aggregated measurement values, serve as dependent features for long-term exposure prediction. Supplementary variables, such as measurement dates, periods, and methods, were employed to filter out invalid observations. A comprehensive data analysis was conducted on all valid data, followed by the extraction of a subset of recent measurements post 2015 for modeling purposes. Metadata for the indoor radon database is extensively documented in Paper VI Appendix A1. The database underpins the regression models for indoor radon concentrations presented in Conf II and the multi-class classification models for intervals described in Paper VI.

Table 2.3. Overview of attributes in the indoor radon database.

Source	Aggregation	Attribute	Variable type
Municipal open database	Building	National property index	Matching key
		Address	Matching key
		Indoor radon measurement dates	Ancillary
		Indoor radon measurement period	Ancillary
		Indoor radon measurement type	Ancillary
		Indoor radon annual average	Independent
		Highest level of indoor radon	Ancillary
		Radioactive concrete	Ancillary
		Floor area	Ancillary
Municipal cadastral register	Building	National property index	Matching key
		Address	Matching key
		County code	Ancillary
		County name	Ancillary
		Municipality code	Ancillary
		Municipality name	Ancillary
		Building category code	Ancillary
		Building type code	Ancillary
		Coordinate longitude	Independent
		Coordinate latitude	Independent
		Geographical adjustment factor	Independent
Energy Performance certificates	Property	National property index	Matching key
		Address	Matching key
		Heated floor area	Independent
		Indoor radon measurement dates	Ancillary
		Indoor radon measurement type	Ancillary
		Indoor radon annual average	Dependent
		Number of floors	Independent
		Number of apartments	Independent
		Number of stairwells	Independent
		Number of basements	Independent
		Ventilation types	Independent
Geological Survey of Sweden databases	Coordinates	Potassium concentration	Independent
		Thorium concentration	Independent
		Uranium concentration	Independent
		Soil type code	Independent
		Soil specification	Ancillary
		Soil collection methods	Ancillary
		Soil collection scale	Ancillary

Investigation of empirical data

A broad examination of building environmental documents was conducted to gather information on past hazardous substance investigations in Sweden. Efforts to find empirical data on asbestos and PCB yielded limited results, as inventory records were stored at individual municipalities, and no data were available on their production plants. However, historical surveys of radioactive concrete and indoor radon at the national scale were found in open databases managed by the Geological Survey of Sweden and some municipalities. This information served as a supplementary data source, enhancing the understanding of radioactive substance presence patterns and their correlation with other factors. Additionally, statistics from these past surveys provided a benchmark for data validation and comparative analysis of results.

From 1929 to 1975, radioactive concrete was extensively used in construction across Sweden. The negative impacts of this material were not recognized until 1980, leading to extensive surveys between 1979 and 1981. These surveys involved vehicle-based gamma radiation measurements across 150 municipalities, as depicted in Figure 2.5. On the mapped districts, the green line indicates the vehicle's route, red circles identify houses with elevated gamma radiation levels due to radioactive concrete, and blue circles mark houses with high gamma radiation that might be caused by radioactive concrete and thus require further investigation. Despite these inventories being outdated due to ongoing changes in the building stock and the lack of follow-up decontamination information, they still provide insight into the prevalence of radioactive concrete in existing buildings. Another search thread led to the discovery of aerial images indicating high gamma radiation, potentially signifying buildings containing radioactive concrete, as shown in Figure 2.6. However, this type of information was sporadically available and challenging to validate on a large scale.

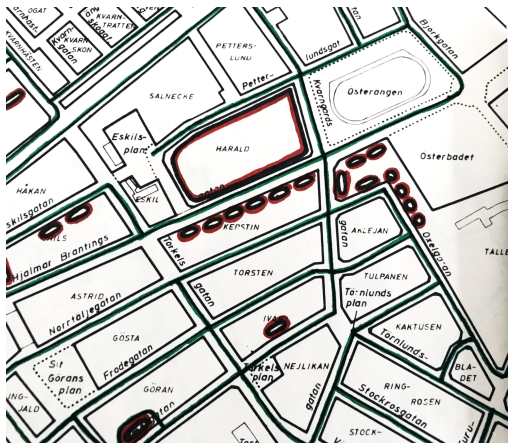


Figure 2.5. Inventory of buildings with radioactive concrete via vehicle measurements of gamma radiation (Geological Survey of Sweden, 1979).

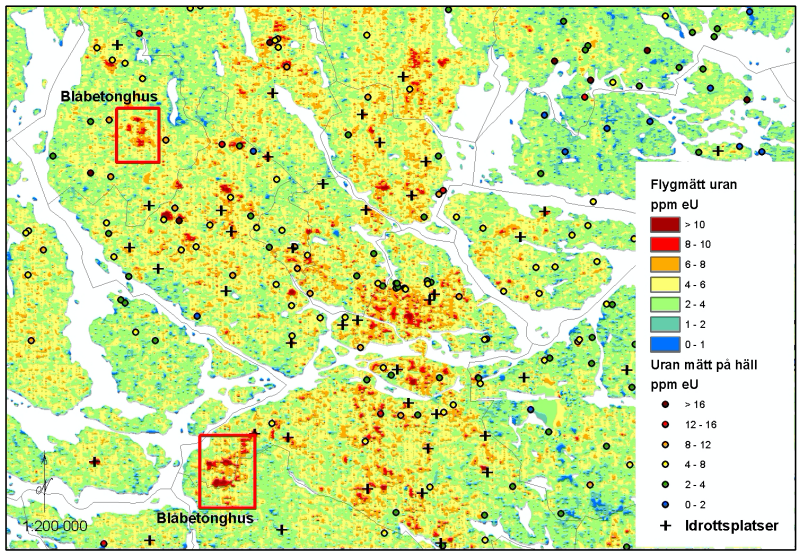


Figure 2.6. Identification of buildings potentially containing radioactive concrete (highlighted in red frames) based on geophysical aerial measurements of uranium gamma radiation conducted by the Geological Survey of Sweden.

In 1983, an indoor radon threshold level of 400 Bq/m^3 was introduced in Sweden, which was further reduced to 200 Bq/m^3 in 2004. Adhering to these regulatory reference levels necessitated the importance of indoor radon measurements, both for monitoring gamma radiation exposure and for implementing remediation in buildings exceeding these thresholds. To discern the impact of geological and geographical factors on indoor radon levels, digital map data regarding radioactive substances and soil types were accessed from the Geological Survey of Sweden, as illustrated in Figure 2.7. Concurrently, indoor radon measurement data were acquired from the open APIs of various municipalities.

Directly inferring the relationship between these variables from the 2D urban-scale radioactive maps proved unfeasible. Therefore, this data was converted into a digital, tabular dataset. This transformed dataset was then combined with additional indoor radon measurements, sourced from EPCs, to facilitate predictive modeling. This integration of diverse data sources allowed for a more comprehensive analysis of indoor radon levels, taking into account both environmental factors and building characteristics.

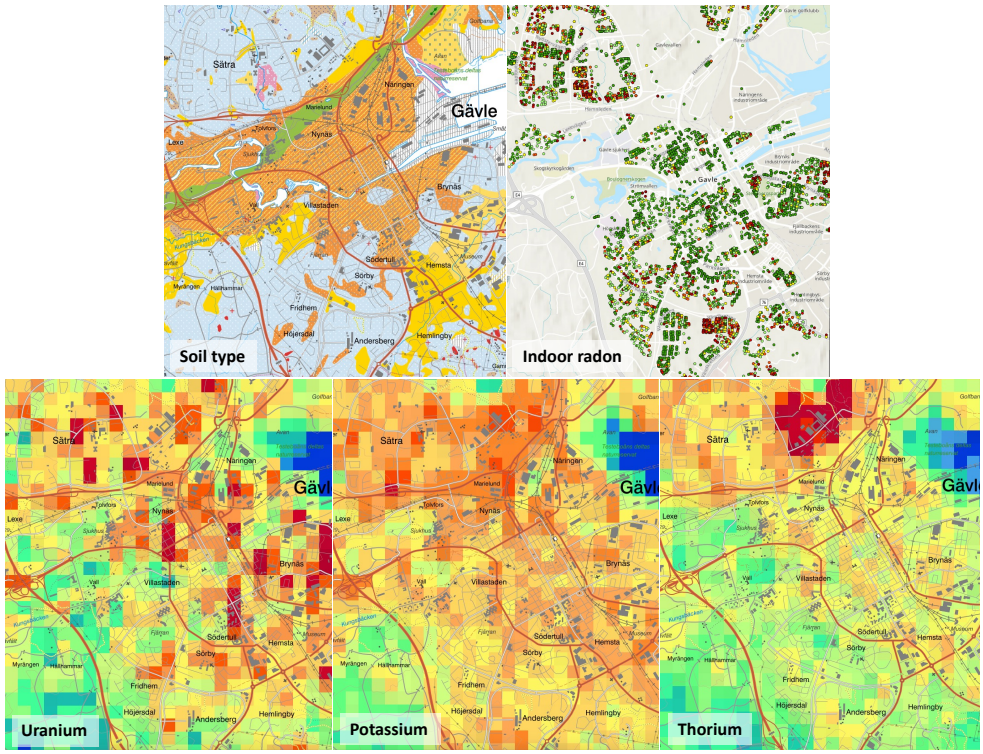


Figure 2.7. Geographical charts of geological data on radioactive substances and indoor radon in Gävle.

2.2.2. Data Preprocessing

Data preprocessing, including data validation and evaluation, data cleaning, missing data imputation, and handling of imbalanced data, was performed on the compiled databases to establish high-quality datasets for subsequent analysis and modeling. Tools such as online spreadsheets, and the Python libraries Numpy (Harris et al., 2020) and scikit-learn (Pedregosa et al., 2011) were employed for these operations.

First and foremost, data validation was conducted to control the consistency in building attributes between building-specific data and building register data. This step was particularly crucial when integrating empirical data sources into established databases for the creation of the hazardous material dataset. Registers sharing the same real estate index were retrieved, and common variables from these data sources, alongside property maps and Google Street Views, were utilized to identify matching buildings. For multi-entry attributes such as construction year, floor area, and building usage, data were consolidated into single entries,

prioritizing accuracy in the following order: inventories, real estate taxation registers, EPCs, and municipal registers.

For the indoor radon dataset, ancillary variables in the measurements were examined to exclude invalid data, such as measurements shorter than two months, those conducted outside the heating season, or prior to the year 2000. In instances where multiple measurements were available for the same property, the most recent data were retained, emphasizing the importance of data timeliness for accurate historical data and updated registers management.

Thereafter, a data assessment matrix was formulated to appraise the quality and quantity of the heterogeneous inventory data. This evaluation took into account multiple dependent variables, building classes, and regional stock compositions. The hazardous material dataset underwent this assessment to pinpoint data subgroups with better quality and more substantial data volume. The following equation, outlined in Papers II-IV, was applied within the data assessment matrix for evaluating the hazardous material dataset.

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

y = Assessment score [0 - 100].
I = Inventory type for weighting observations. I = 1 if is report (r), I = 0.75 if is protocol (p), I = 0.5 if is control plan (c), and I = 0.25 if is demolition plan (d).
n = The number of observations in the subgroup [0 < n].
N = The number of observations in the entire dataset.
K = Weight based on data size. K = 1 if n >= 400, K = 0.75 if 300 =< n < 400, K = 0.5 if 200 =< n < 300, K = 0.25 if 100 =< n < 200, K = 0 if n < 100.

For each subgroup of hazardous materials and building classes, a larger number of detailed inventory implies a higher assessment score and a greater modeling potential. In the case of continuous indoor radon measurements, a descriptive analysis was conducted to assess data count, value range, and mean values across building classes, aiding in determining the distribution and variations between different data subgroups. Data cleaning was performed to selectively refining data subsets of interest and eliminating incorrect, corrupted, or duplicated entries. Given that various hazardous substances were commonly used in construction between 1930 and 1980, buildings from this era within the Swedish building stock were a primary focus. This specific subset was extracted from both the hazardous materials and indoor radon databases for outlier detection and removal. The Interquartile Range (IQR) method, which is particularly effective for non-Gaussian distributions by calculating cutoffs at 1.5 times the IQR from both the 75th and 25th percentiles, was utilized in Paper VI and Conf II to identify and exclude anomalous data for predictive variables. Subsequently, the numerical data were standardized, and categorical data were encoded for further processing.

Observations containing large numbers of missing data in the datasets were either removed or imputed based on the types of missing data, Missing Completely at Random, Missing at Random, and Missing Not at Random. The Python library missingno was employed to visually represent the presence and extent of missing data in matrix and bar formats. Variables with missing values exceeding 30% were excluded from the feature set to minimize their impact on prediction results. This process was tailored to individual data subgroups due to considerable variations in variable availability among building classes and regions. Continuous variables identified as Missing Completely at Random were imputed using the *k*-NN algorithm in Paper IV, where the average of *k* nearest neighbors was calculated for imputation uniformly or weighted by distance. Correlations between dependent and independent variables guided the feature selection in multivariate imputation.

Both the hazardous material dataset and the indoor radon dataset showed skewed class distributions, characterized by a low incidence of positive hazardous material detections and few observations with elevated indoor radon levels. This imbalance led to classifier bias towards the majority class, often resulting in misclassification of the minority class. To address this, various techniques were employed, including resampling, data augmentation, algorithm adjustments, selecting appropriate evaluation metrics, and threshold moving. Initially, oversampling of the minority class, either through replication or using the Synthetic Minority Oversampling Technique (SMOTE) to create synthetic instances, was applied. Algorithm adjustments were then made by incorporating sample weights based on inverse class frequency. Evaluation metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC) and the F1 score (the harmonic mean of precision and recall) were chosen over accuracy to assess imbalanced classification more effectively. Finally, optimal threshold values were determined through threshold shifting in ROC curves and Precision-Recall curves to maximize AUC and F1 scores.

2.2.3. Explorative Data Analysis

Data analytics were conducted on the cleaned datasets to establish a preliminary understanding of the dataset structures and the interrelationships among variables. Utilizing Python's statistical visualization libraries, Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021), sample distribution and correlation analysis for both the hazardous material dataset and the indoor radon dataset were carried out. These analyses became a foundation for subsequent algorithm selection and variable transformation. Thereafter, feature selection was performed with scikit-learn to pinpoint critical features for statistical and machine learning modeling.

The analysis of sample distribution, which reflects the probability distribution of statistics from numerous samples drawn from a specific population, was pivotal in identifying potential skewness in the dataset. This involved estimating standard

deviations, confidence intervals, and the upper and lower limits. Graphical representations of numerical variables, such as construction year, floor area, number of floors, and indoor radon measurements, were created in various formats including histograms, scatterplots, boxplots, violin plots, and strip plots. Boxplots were used for comparing the summary statistics regarding the spread of mean, median, minimum, maximum, and outlier values across different data subgroups. Violin plots complemented this by showcasing the kernel density distribution of each variable, providing insights into the shape of the distribution. The density distribution or probability distribution was then normalized to depict the frequency distribution of building parameters across both training and prediction sets, aiding in assessing how representative the building samples were of municipal or regional building stock. Lastly, strip plots, which are essentially categorical scatterplots with jitter, were employed to enhance the visualization provided by violin plots. These plots were instrumental in representing the underlying distribution of the data, offering a more nuanced view of the relationships and patterns within the dataset.

Correlation analysis, also known as bivariate analysis, was implemented to explore the relationships between dependent and independent variables and to quantify these relationships in terms of correlation coefficients. The hazardous material and indoor radon datasets, which comprise a mix of data types, were analyzed using Pearson correlation matrices to identify relationships between pairs of variables. These relationships were visually represented in heatmaps. Clustergrams, which combine heatmaps with dendrograms, facilitated hierarchical cluster analysis and provided a visual means to understand the patterned clustering of variables. In addition to this, the p-value (the level of marginal significance within a hypothesis test) for each pairwise relationship was computed to ascertain which variables significantly enhanced model fit in statistical terms.

Feature selection was integrated into a model-based pipeline to minimize the number of input variables. This reduction not only lowers computational costs but can also enhance model performance. Various techniques were employed in the thesis based on their unique strengths, such as f-test in the analysis of variance (ANOVA), selectKbest, selectPercentile, and recursive feature elimination (RFE). The F statistic in ANOVA measures the joint effect of all variables, helping to determine if the variance between the means of two populations is significant, and is typically used in conjunction with p-values. The selectKbest and selectPercentile algorithms, which operate based on f-values, return either the top k highest scores or a certain percentile of the highest scores, respectively. In contrast, RFE iteratively identifies the best feature set by gradually removing the least important features until the desired number of features remains. The selection and number of features are closely linked to model performance. Therefore, datasets derived from different data preprocessing methods were examined during model training to identify optimal combinations.

2.3. Quantitative Approaches

Data-driven approaches encompassing statistical modeling, machine learning, and neural networks. These were employed to develop predictive models for hazardous substances. Given the non-linear relationships and intricate interdependencies among variables, non-parametric algorithms were selected for predictive modeling. These algorithms do not presuppose a specific model structure; rather, they adapt to the form of the data. Figure 2.8 presents the mapping of algorithms, evaluation metrics, and model explainability techniques employed in the thesis. Papers III and IV explored numerous supervised learning classifiers to predict asbestos and PCB-containing materials in buildings, while Paper V utilized a statistical approach for estimating the presence of radioactive concrete. The prediction of indoor radon concentrations was addressed in Conf II, which compared the performance and feature selection between statistical and machine learning models. Paper VI advanced this line of inquiry by predicting indoor radon levels using both machine learning and neural network modeling.

The subsequent section provides a detailed overview of each component of the modeling pipeline. This includes model training, where algorithms learn the best combination of weights and bias to minimize a loss function; hyperparameter tuning, which involves optimizing the settings within each model to improve performance; model evaluation, where the effectiveness of a model is assessed; and output interpretation, which involves making sense of the model's predictive results. This comprehensive approach ensures that the predictive models are not only accurate but also interpretable and relevant to the needs of the study.

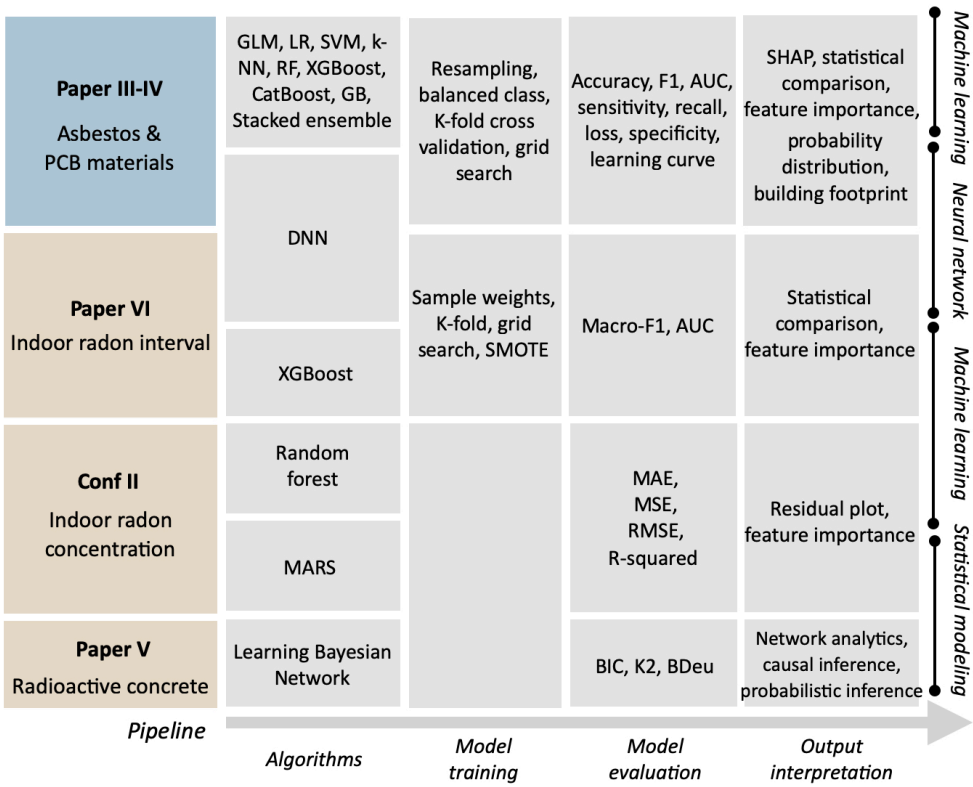


Figure 2.8. Mapping of predictive modeling approaches used in the papers.

The hazardous material and the indoor radon datasets were partitioned into training and validation sets with comparable proportions of each label. The training data then underwent a series of steps in the machine learning pipeline – model training, hyperparameter tuning, model evaluation, and output interpretation. K-fold cross-validation was employed on the training set, which averages model performance over k iterations. Given the class imbalance present in both datasets, the Synthetic Minority Oversampling Technique (SMOTE) was used to augment minority classes such as positive detection records and high indoor radon intervals. Additionally, sample weights within algorithms were adjusted to address this imbalance. Optimal threshold moving was also explored to minimize false negatives and maximize AUC and (macro-)F1 in the ROC curve and Precision-Recall curve. Thereafter, a random grid search was conducted to evaluate various combinations of hyperparameters, fine-tuning the model until the point of minimum loss. The lead models' performance in both cross-validation and the validation set (approximately 20% of the dataset) were assessed using the F1 and AUC metrics for classification tasks, and Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared for regression tasks. The influence of data

size on model performance was evaluated by analyzing learning curves for both training and validation.

To untangle the gray-box models and interpret the prediction outputs, explanation plotting functions such as variable importance, SHapley Additive exPlanations (SHAP), and partial dependence (PD) plots were utilized. Variable importance, calculated from the gains in loss functions during the construction of tree-based algorithms, highlighted the relative influence of each feature. SHAP values quantified the contribution and magnitude of feature impact on both local (individual observation level) and global (entire dataset level) predictions. PD plots revealed the marginal effect of a single feature on the dependent variable across various models, assuming independence between features. By measuring the change in the mean response, the average impact of variables on the model's predictions was ascertained. Finally, the developed models were applied to the prediction set. The outcomes of these predictions were then analyzed using statistical description and geospatial visualization tools, providing a comprehensive understanding of the model's applicability and effectiveness.

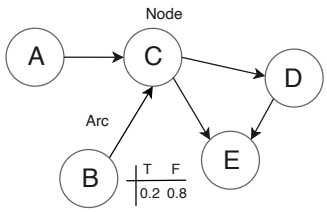
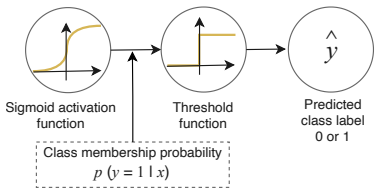
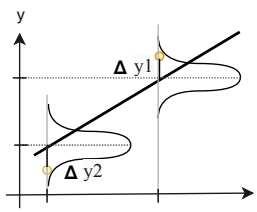
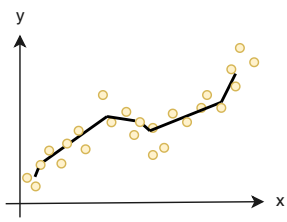
2.3.1. Statistical Modeling

Statistical modeling serves as an alternative inference approach to machine learning, for approximating reality and deriving inferences. Using statistical hypothesis tests and statistical estimators, the mathematical relationship between a dependent variable and one or more independent variables can be mapped into functions. Statistical models are non-deterministic and stochastic, representing probability distributions rather than specific values (James et al., 2023), and explaining relationships between variables (Hastie et al., 2016). Hence, these models are versatile, and their prediction results are straightforward to interpret, supporting deductive-, inductive-, and abductive reasoning. Statistic modeling is particularly prevalent in the AEC sector, favored for its high interpretability, ease of implementation, and established methodology (Wei et al., 2019). They have shown commendable performance in handling incomplete data, as evidenced in the literature. Examples include the ontology-based approach for predicting Asbestos-Containing Materials (ACM) in buildings based on product timelines (Mecharnia et al., 2019), and a multicriteria analysis of existing building stocks using Bayesian Networks (Carbonari et al., 2019). Nevertheless, statistical models have shortcomings in terms of complexity and their capacity to manage real-time, high-dimensional, mixed-source datasets common in construction settings, especially when dealing with missing or noisy data (Yan et al., 2020).

In this thesis, statistical models were developed to extract insights from historical data, fit the stochastic nature of the compiled data, and construct predictive models for statistical inference. Python libraries pgmpy (Ankan & Panda, 2015) and py-earth (Friedman, 1991; Rudy, 2013) were used for modeling Learning Bayesian

Networks (BN) in Paper V and Multivariate Adaptive Regression Splines (MARS) in Conf II, respectively. Logistic Regression (LR) and Generalized Linear Model (GLM) were built in Papers III and IV via Python libraries scikit-learn and H2O. A comprehensive overview of the architecture, strengths, and limitations of each statistical model is presented in Table 2.4 below.

Table 2.4. Overview of statistical modeling algorithms used in the thesis (adapted from James et al., 2023).

Algorithm	Architecture	Strength & limitations
Learning Bayesian Network		+ Graphical and interpretable + Model interdependencies + Account for uncertainty - Limited to discrete variables - Undetermined prior selection - Computationally demanding
Logistic regression		+ Classification + Simple implementation + Perform well in low-dimensional or small datasets - Multi-collinearity - Limited to binary classification
Generalized linear model		+ Classification & regression + Stepwise forward or backward variable selection - Sensitive to outliers - Multi-collinearity
Multivariate Adaptive Regression Splines		+ Regression + Flexible implementation + Robust to outliers + Easy interpretation - Susceptible to overfitting - Cannot handle missing values

Bayesian Network (BN), a framework for representing and reasoning under conditions of uncertainty (Cheng et al., 2002), shows promising results in estimating the presence of hazardous materials, i.e., radioactive concrete in buildings. Algorithms that utilize information-theoretic analysis for learning BN structure from data were employed to estimate the probability of radioactive concrete in buildings via structure learning and parameter learning (Cheng et al., 2002). Eight

variables suspected to be associated with the presence of radioactive concrete were discretized into three to five intervals using binning techniques, such as construction year, floor area, building class, the average distance to historical radioactive concrete manufacturing plants, number of floors, stairwells, and apartments. Structural learning algorithms were then used to construct the underlying graph skeleton, including the mapping of factor dependencies and the determination of edge direction in the graph. Subsequently, the generated directed acyclic graphs (DAGs) underwent evaluation using scoring functions such as K2, BDeu (Bayesian Dirichlet equivalent uniform), and BIC (Bayesian Information Criterion). The DAGs with the highest scores were selected for further parameter learning, where we computed the conditional probability distribution (CPD) using Bayesian Parameter Estimation.

After configuring the BN models, causal inference and network analytics, i.e., predictive, diagnosis, and sensitivity analysis, were performed to enhance the understanding of the networks' behavior. Predictive analysis assessed the models' performance in predicting the presence of radioactive concrete, varying the number of evidence inputs to reflect data availability. Diagnostic analysis focused on identifying key factors in detecting radioactive concrete by examining changes in posterior probabilities. Sensitivity analysis involved rebinning variables, particularly building classes, to pinpoint those contributing significantly to output variation. Thereafter, the Bayesian network models were transformed into causal network models to identify causal dependencies between factors and applied the backdoor adjustment method to iterate models with and without do-adjustment.

Parametric models, specifically logistic regression (LR) and generalized linear models (GLM), were utilized to classify materials containing asbestos and PCB. LR, as described by Raschka & Mirjalili (2019). They fit data using predefined functions with fixed parameters and are suitable for estimating binary outcomes' probabilities. The LR models assume a linear relationship between dependent and independent variables, serving as a baseline against more sophisticated statistical and machine learning algorithms during model development. Conversely, the GLM models extend ordinary linear regression to accommodate non-linear, categorical, or continuous data. This extension, noted by McCullagh & Nelder (1983), allows the variance of each observation to be modeled as a function of its predicted value. By employing appropriate link functions – normal density for normal distribution, logit or sigmoid for binomial distribution, and log for Poisson distribution – the linear association of dependent variables is facilitated. The performance of developed classifiers was then evaluated with the confusion matrix and the significance of variables was assessed through coefficient magnitudes.

For modeling indoor radon concentrations, non-parametric multivariate adaptive regression splines (MARS) models were employed, given the dataset's non-Gaussian and non-linear characteristics. The MARS algorithm adeptly fits the dataset with either piecewise linear or cubic splines, utilizing hinge functions for automatic adjustment. This method, as articulated by Kartal Koc & Bozdogan

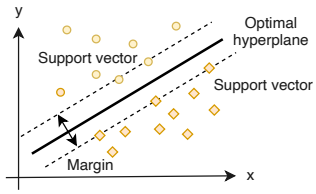
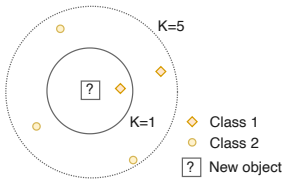
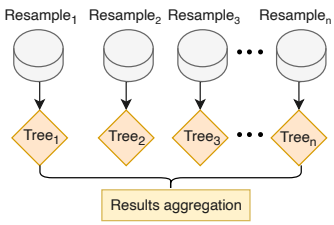
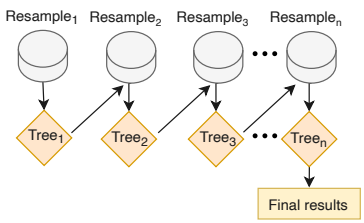
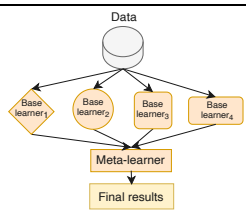
(2015), is represented through equations encompassing intercepts, coefficients, hinges, and features. The MARS models' performance was evaluated using metrics including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R-squared). These metrics were instrumental in quantifying deviations between predicted and actual values across the training and validation sets. Furthermore, the models' the the model fit to the data and the assumptions underlying linear regression, quantified as the differences between the observed values and the values predicted by a model, was evaluated using residual plots. Then Q-Q plots were created to compare two probability distributions by plotting their quantile against each other, facilitating data distribution comparison and data skewness assessment.

2.3.2. Machine Learning

Machine learning approaches are differentiated based on the availability of labels in the data, broadly classified into supervised learning, unsupervised learning, and reinforcement learning, as categorized by (Raschka & Mirjalili, 2019). Supervised learning involves fitting models to labeled datasets to generate predictive outcomes in regression and classification. In contrast, unsupervised learning focuses on representation learning without predefined labels, aiming to uncover hidden patterns in data through methods such as clustering and association. Reinforcement learning, meanwhile, involves an agent learning to make decisions in a dynamic environment, balancing between exploration and exploitation within a framework of actions and rewards.

This thesis delves into various non-parametric algorithms, significant for their lack of predefined assumptions about data mapping functions. These algorithms have been explored and compared across diverse problem settings. They are primarily categorized into model types such as distance-based algorithms, including support vector machines (SVM) and *k*-nearest neighbors (*k*-NN), as well as decision tree-based models including random forest (RF), gradient boosting (GB), extreme gradient boosting (XGBoost), categorical boosting (Catboost), and stacked ensemble. These categories are detailed in Table 2.5. The focus here is on supervised learning models, developed to benchmark their performance in predicting contamination in buildings at different lifecycle stages, including maintenance, retrofit, and end-of-life. For model development, Python libraries scikit-learn and H2O were utilized, as detailed in Papers III, IV, VI, and Conf II.

Table 2.5. Overview of machine learning algorithms used in the thesis (adapted from Raschka & Mirjalili, 2019).

Algorithm	Architecture	Strength & limitations
Support vector machine		+ Classification & regression + Perform well in small datasets + Flexible implementation - Susceptible to dimensionality - Cannot handle missing values - Sensitive to outliers
k-Nearest Neighbors		+ Classification & regression + Flexible implementation + No training required - Susceptible to dimensionality - Sensitive to k & distance metric - Sensitive to outliers
Random forest (Distributed random forest/Extremely randomized trees)		+ Classification & regression + Handle high dimensional data + Handle missing values + Reduced variance + Parallel processing - Computationally expensive - Susceptible to overfitting
Gradient boosting/ Extreme gradient boosting/ Categorical boosting		+ Classification & regression + High accuracy and robust + Handle high dimensional data + Handle missing values + Reduced bias and overfitting + No regularization required - Require extensive tuning
Stacked ensemble		+ Classification & regression + Handle high dimensional data + High accuracy and robust - Increase the risk of overfitting - Complex implementation

The kernel SVM operates by maximizing the margin space around the decision boundary, delineating an optimal hyperplane for handling non-linear data, as outlined by Raschka & Mirjalili (2019). In this study, the radial basis function kernel was employed, with careful tuning of the gamma parameter (defining kernel width) and the C parameter (regulating the trade-off between maximizing the margin and

minimizing misclassification errors). Concurrently, the *k*-nearest neighbors (*k*-NN) algorithm, a distance-based method, was utilized. It classifies instances based on a majority vote from the *k* nearest neighbors. The choice of *k* is important, as a smaller k might increase sensitivity to noise, whereas a larger *k* could lead to underfitting. The kernel SVM and *k*-NN were selected for their adaptability in small dataset scenarios and efficiency in achieving high performance with limited computational resources. However, these models also have limitations, including sensitivity to dimensionality and outliers, and inability to produce insights on feature importance.

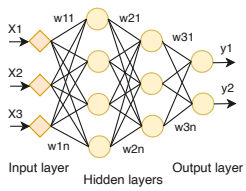
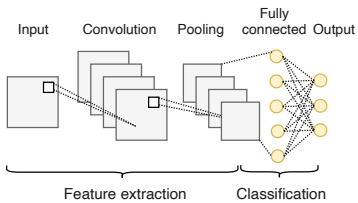
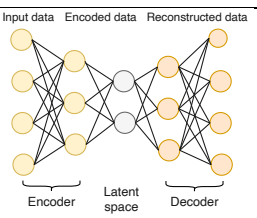
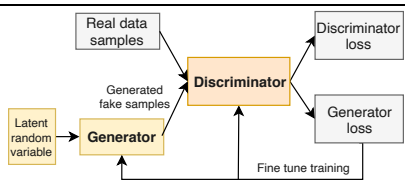
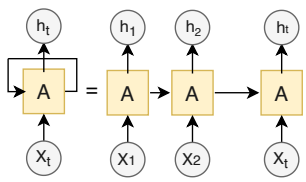
Decision-tree-based models were employed for their robustness and high performance in classification and regression tasks. These scale-invariant tree ensemble models combine outputs from base learners to train a meta-learner, aiming to enhance prediction accuracy while reducing errors. This is achieved through various methods: bagging for fully-developed decision trees, boosting for shallow trees, and hybrid techniques such as stacking, blending, voting, and cascading. Specifically, Random Forest (RF) utilizes bagging to combine multiple predictions from resampled base learners, while Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost) leverage boosting to sequentially transfer predictions from one learner to the next, focusing on correcting misclassifications made by predecessors. XGBoost and CatBoost are particularly adept at handling numerical and categorical data, respectively, employing gradient boosting to iteratively generate weak learners in a manner that minimizes loss and reduces the likelihood of overfitting. Recognizing the strengths and weaknesses of each method, four types of tree ensemble learning were trained in parallel as sub-models and then integrated into a stacked meta-learner to determine the optimal combination of model contributions.

2.3.3. Artificial Neural Network

The experimental component of the thesis delves into artificial neural networks (ANNs) and deep neural networks (DNNs), integral to the machine learning (ML) field, facilitating both supervised and unsupervised representation learning. Given the limited size and low dimensionality of the hazardous material and indoor radon datasets, neural networks with simpler architectures were evaluated for their performance relative to traditional ML models. ANNs, as noted by Goodfellow et al. (2016), are capable of automatic feature learning and handling missing data, though they necessitate categorical embedding during data preprocessing, contrasting with distance-based or tree-based models. Multilayer perceptrons (MLPs) were selected as the primary neural network type, considering their suitability for tabular data without time dependencies, such as inspection and measurement records. These MLPs were used to estimate the probabilities of hazardous materials and indoor radon concentrations. The development of these multilayer, feed-forward neural network models utilized Keras, a Python interface

for TensorFlow, and the H2O library. These models were trained employing stochastic gradient descent with back-propagation, as detailed in Papers IV and VI. Additionally, other deep learning applications were suitable for other purposes: convolutional neural networks (CNNs) for image recognition and classification, autoencoders and generative adversarial networks (GANs) for dimensionality reduction and information retrieval, and recurrent neural networks (RNNs) for modeling serial data with consideration for short and long-term temporal dependencies. These applications are outlined in Table 2.6.

Table 2.6. Overview of neural network algorithms (adapted from Raschka & Mirjalili, 2019).

Algorithm	Architecture	Strength & limitations
Multilayer perceptrons		+ Classification & regression + Handle missing values + Parallel processing + Automatic feature learning - Require extensive training - Difficult model interpretation
Convolutional neural network		+ Image recognition& classification + High performance + Automated feature extraction + Robust to noise + Support transfer learning - Computationally expensive - Difficult with small datasets
Autoencoder		+ Dimension reduction + Automated feature extraction + Anomaly detection - Computationally expensive - Difficult latent space interpretation
Generative adversarial nets		+ Data augmentation + Semi-supervised learning + Anomaly detection + Missing data imputation - Require extensive training - Risk of model collapse
Recurrent neural network		+ Sequential and time-series data + A wide range of applications - Require extensive training - Prone to exploding - Prone to gradient vanishing

Hyperparameter tuning was performed to search for the optimal combination of layers and neurons, activation functions, optimization methods, learning rates, dropout rates, training epochs, and batch size for the neural network models. The choice of output and loss functions was tailored to the nature of the classification task: sigmoid activation with binary cross-entropy for binary classification, and softmax activation with categorical cross-entropy for multi-class classification. To address the issue of imbalanced class distribution in the dataset, sample weights were calculated and integrated into the weighted metric.

Further, the training history of lead models was examined, where loss was plotted against training epochs or the number of trees. This analysis was conducted across the train, cross-validation, and validation sets to gain insights into the models' learning and generalization behaviors, particularly focusing on their capability to fit data and maintain representativeness. The performance of these models was evaluated using the same metrics as the preceding models, specifically macro-F1 score and area under the curve (AUC). Apart from these evaluations, an in-depth analysis of influential features in the ANN models was performed. This analysis was grounded in the Gedeon method (Gedeon, 1997), which considers the weights connecting input features to the first two hidden layers, thereby providing insights into the features' contributions to predictions.

2.3.4. Predictive Analytics

The predictive component of this research involved extracting features from the Swedish building stock for buildings that were not inspected or measured, specifically to estimate the probability of containing hazardous substances. In Paper IV, this approach yielded predictions regarding the presence of asbestos and PCB-containing materials, with results including both predicted labels and probabilities for buildings in Stockholm, Gothenburg, Malmö, and Kiruna, as depicted in Figure 2.9. The distribution of these probabilities, with respect to specific asbestos and PCB-containing materials, was analyzed across various building categories and construction years. Subsequently, OSMnx—a Python library tailored for street network data collection and analysis—was employed to obtain building footprint data from OpenStreetMap (Boeing, 2017). This data facilitated the creation of choropleth maps, which visualized the predicted probabilities of asbestos and PCB materials in building clusters, with a focus on municipality-owned residential buildings in Stockholm.

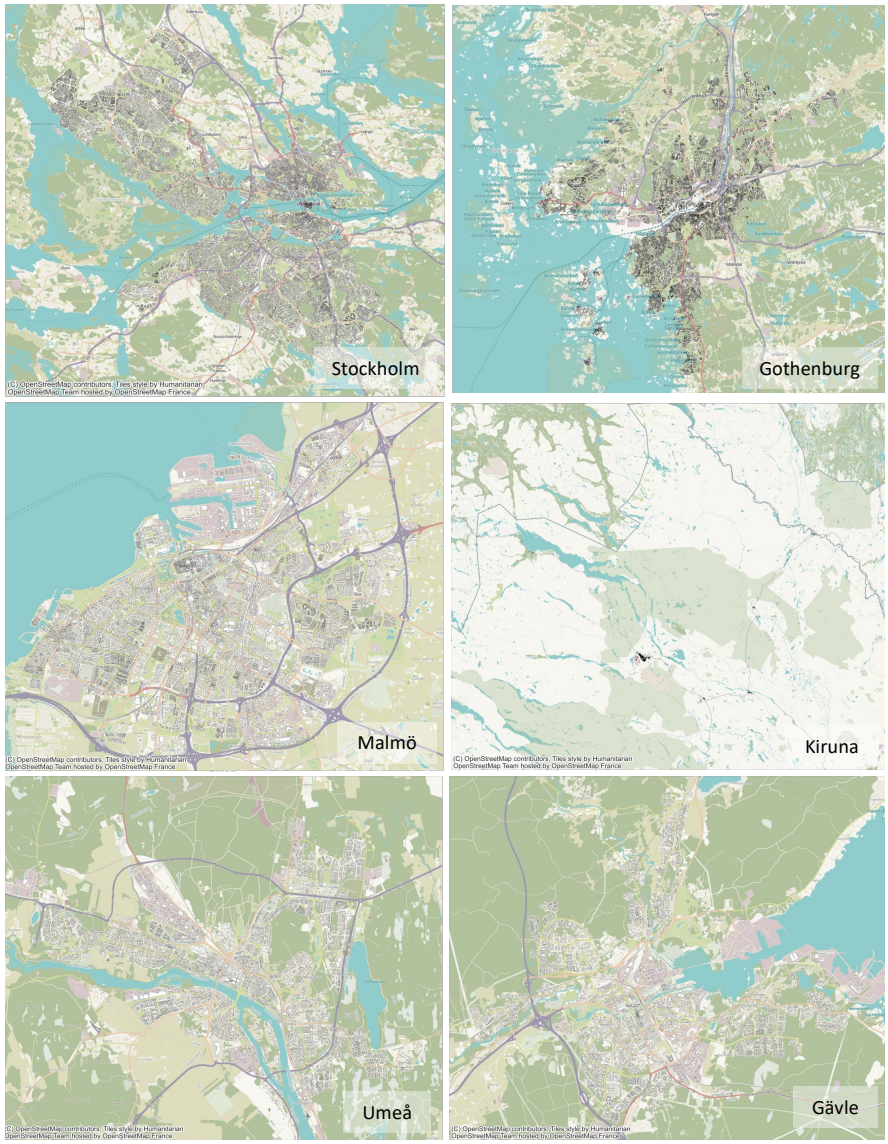


Figure 2.9. Building footprint maps of the studied municipalities (OpenStreetMap, 2023).

Paper V extended this methodology to the predictive analysis of radioactive concrete in five municipalities: Stockholm, Gothenburg, Malmö, Gävle, and Umeå. This analysis utilized probabilistic inference based on Bayesian hierarchical modeling, integrating three base models to produce joint and disjoint probability distributions. These distributions were conditioned on key variables such as building class, construction year, presence of a basement, floor area, and proximity to known

historical radioactive concrete plants. By interfacing with local building registers, the models facilitated the estimation of probabilities for individual buildings, considering the combined influence of multiple variables.

The prediction of indoor radon levels was specifically focused on the Swedish metropolitan areas of Stockholm, Gothenburg, and Malmö. These regions were selected due to the high volume of indoor radon measurements available, with their geographical representation detailed in Figure 2.10. In Paper VI, statistical summaries of historical indoor radon measurements were retrieved to allow the comparison with the predicted measured and predicted non-measured buildings within these regions. Such a comparative approach enabled a thorough evaluation of the model's sensitivity in predicting indoor radon levels on a regional scale and the reliability of the resultant predictions. The outcomes of these predictions were methodically detailed, showcasing the distribution of building shares with varying indoor radon levels by region and building category.

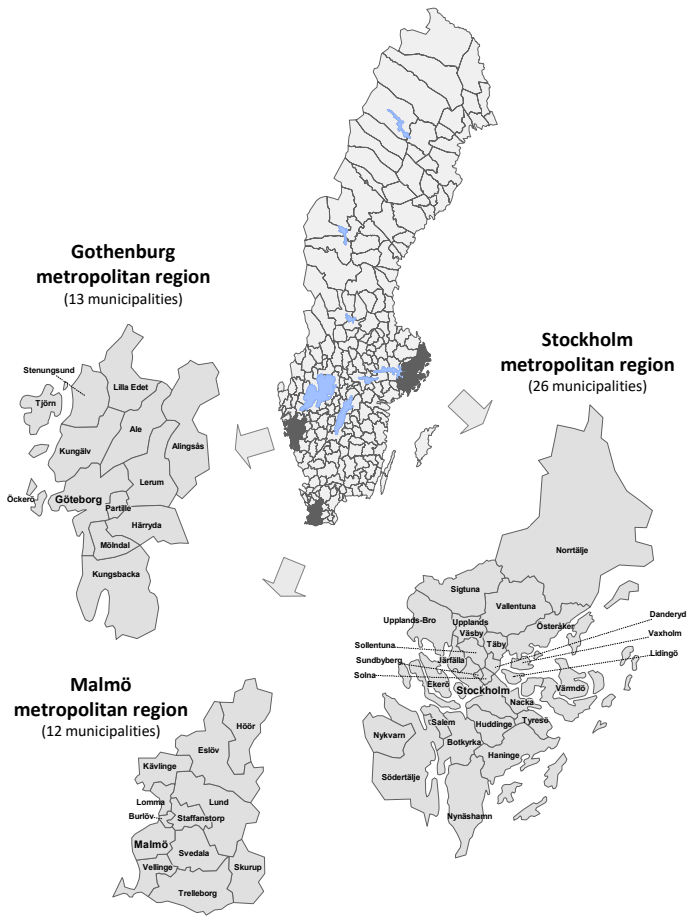


Figure 2.10. Swedish metropolitan areas (SCB, 2005).

3. Research Findings

The chapter is structured according to the research layout connecting research questions, research scope, and articles, as shown in Figure 3.1. Section 3.1 explores data-driven approaches in managing in situ hazardous materials and examines the CDW industry, incorporating insights from a stakeholder workshop and an expert interview. Section 3.2 outlines the workflow for curating and processing data from pre-demolition audits, essential for assessing hazardous material records quality and quantity (RQ1). Sections 3.3 and 3.4 delve into data analysis and predictive modeling for hazardous substances, with Section 3.3 focusing on asbestos and PCB (blue) and Section 3.4 on radioactive concrete and indoor radon (brown). These sections form the basis for a statistical modeling and machine learning pipeline (RQ2) and explore model application in estimating probabilities of hazardous substance in the Swedish building stock (RQ3).

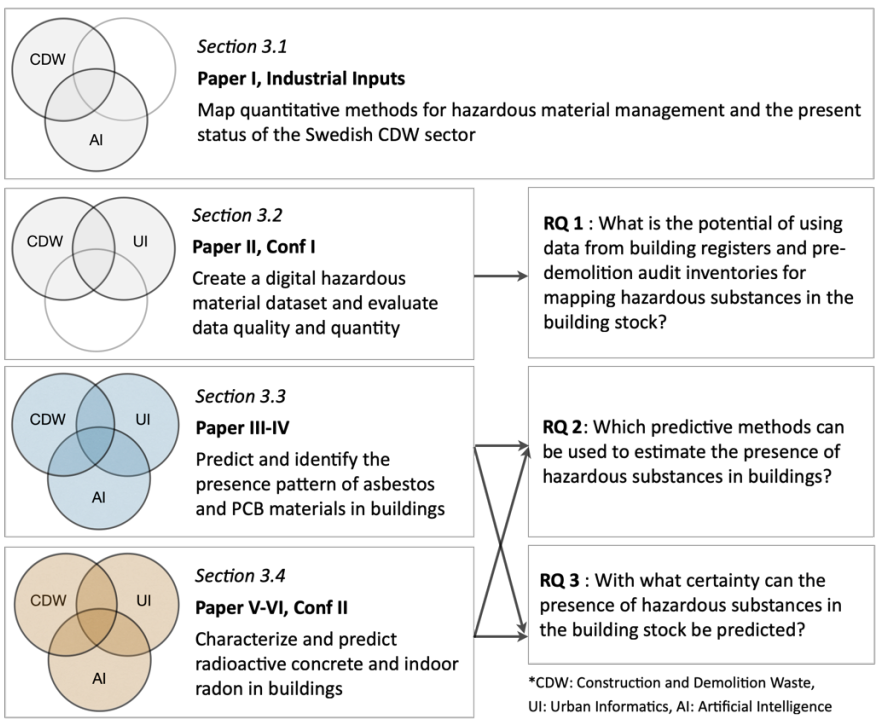


Figure 3.1. Research layout.

3.1. State-of-the-art

The section offers a comprehensive overview of the present state of research and industrial practices related to in situ hazardous material management. Section 3.1.1 details various data sources and digital tools used for identifying hazardous materials at multiple scales, recognizing waste, and facilitating onsite sorting. This includes a detailed examination of technologies and methodologies employed in detecting and managing hazardous materials at different scales. Subsection 3.1.2 presents insights gleaned from engaging dialogues with stakeholders in the Swedish construction sector. These discussions focus on current practices in hazardous material identification and decontamination processes, providing a practical perspective on the topic. The cross-comparison allows the thesis to identify and establish clear research directions, particularly regarding the need for and potential benefits of digital applications in hazardous material management. This approach ensures that the research is grounded in both theoretical understanding and practical realities, making it relevant and applicable to ongoing challenges and advancements in the field of construction and demolition waste management.

3.1.1. Research Front of Hazardous Material Management

The science mapping conducted in Paper I revealed a complex interplay among various research domains, with significant publication increases over years in Environmental Sciences and Ecology (34%), Public, Environmental, and Occupational Health (23%), and Engineering (18%) between 1990 and 2020, focusing on hazardous building material management. A predominant theme in earlier studies was the measurement of exposure and remediation of risks associated with specific contaminants, particularly asbestos and PCB materials. This indicates a shift in the perception of hazardous materials, transitioning from an occupational to a public health hazard, particularly in relation to existing building stocks. Surprisingly, the literature on radioactive concrete was scarce, with only a marginal percentage (approximately 4%) of references discussing in situ hazardous materials in buildings. The few existing studies, primarily published in journals related to Construction and Building Technology, concentrated on sampling, monitoring, mitigation, remediation measures, and risk management in disaster scenarios. The paucity of literature shows a significant gap in systematic perspectives on managing the risks of hazardous substances in existing buildings.

Furthermore, the co-word analysis identified “discipline silos,” particularly in terms of the frequency and extent of comprehensive hazardous material studies. Two distinct conceptual clusters were recognized: one focusing on asbestos-related diseases for workers, and the other on indoor exposure to PCB-contaminated materials. Word dynamic analysis, tracking the cumulated occurrences of keywords,

showed a marked increase in attention to asbestos and PCB exposure in buildings since 2013. Moreover, the intellectual structure's direct citation network revealed an evolving research paradigm over the past three decades. Initial studies in the 1990s focused on asbestos occurrence and decontamination, shifting towards asbestos-related diseases in subsequent years. Since the 2000s, the scope broadened to encompass the relationship between pollution and health, paralleled by research into mapping asbestos roof coverings and urban-scale inventories. Studies on PCB sources and emissions also gained traction during this period, evidenced by a more interconnected citation network. With the increasing adoption of circular economy principles in the construction sector, ensuring the safety of material disassembly, reuse, and recycling has become a pivotal concern. A surge in related research was observed in the 2010s, with several studies focusing on characterizing hazardous materials in existing buildings and construction and demolition waste. As depicted in Figure 3.2, these studies span over various scales and purposes, employing data-driven approaches for hazardous material management.

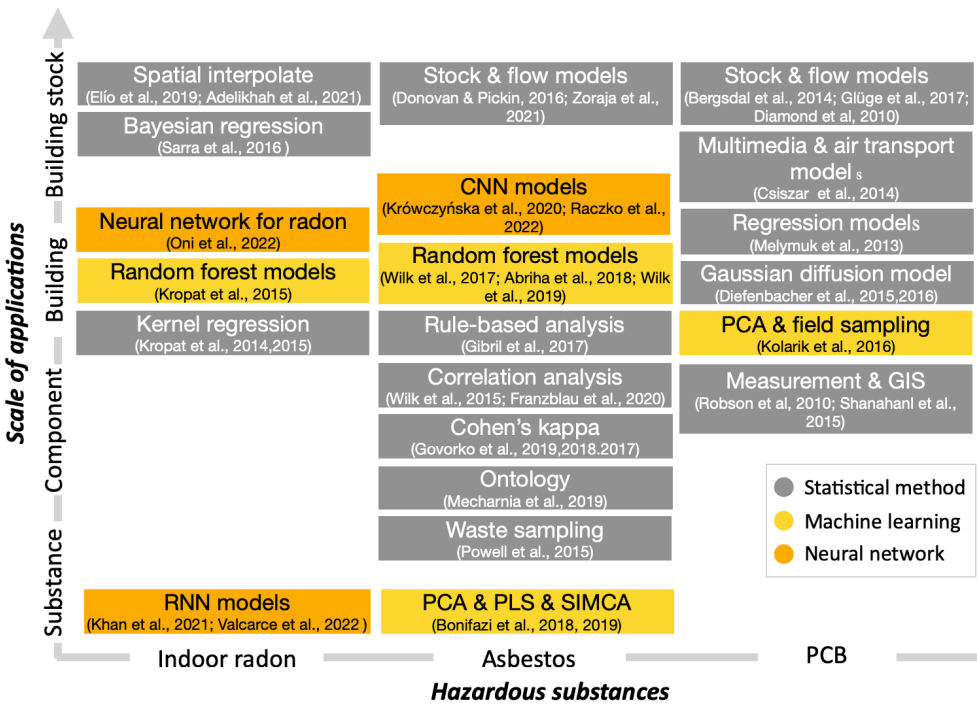


Figure 3.2. State-of-the-art applications for hazardous material management organized by scales and purposes (adapted from Paper I with additional literature on indoor radon and PCB).

Among these approaches, analytical or statistical methods (highlighted in gray) emerged as the most prevalent in addressing different aspects and scales of

hazardous material management. For example, stock and flow models were utilized to estimate the lifecycle of asbestos products (Donovan & Pickin, 2016; Zoraja et al., 2021) and PCB (Bergsdal et al., 2014; Diamond et al., 2010; Glüge et al., 2017) in national building stocks. The mapping of asbestos-cement roofing and the factors influencing their quantities were explored through rule-based analysis for image classification (Gibril et al., 2017) and correlation studies at a regional scale (Wilk et al., 2015). Detailed descriptive analyses of specific asbestos components in buildings were conducted using various methodologies, including Cohen's kappa statistics (Govorko et al., 2017, 2018, 2019), ontology and rule-based methods (Mecharnia et al., 2019), Pearson correlation (Franzblau et al., 2020), and waste sampling during source separation (Powell et al., 2015).

In addition, the diffusion of PCB products at the urban scale was estimated using multimedia and air transport models (Csiszar et al., 2014), regression models (Melymuk et al., 2013), and Gaussian diffusion models (Diefenbacher et al., 2015, 2016). At the substance level, the bottom-up source inventory of specific PCB components was mapped using sampling and GIS tools (Robson et al., 2010; Shanahan et al., 2015). Concurrently, indoor radon maps were developed on continental (Elío et al., 2019) or national scales (Adelikhah et al., 2021; Kropat, et al., 2015a; Sarra et al., 2016) employing spatial interpolation techniques, quantile Bayesian regression, and kernel regression.

The systematic literature review showed a limited number of studies integrating supervised or unsupervised machine learning methods (highlighted in yellow) with empirical data for hazardous substance quantification. Prominent examples include the application of the random forest algorithm for predictive mapping of asbestos-cement roofing, utilizing remote sensing and physical inventory data (Abriha et al., 2018; Wilk et al., 2017, 2019), and indoor radon concentration prediction based on the automated classification of lithological units (Kropat et al., 2015b). Additionally, non-destructive methods for asbestos fiber detection were proposed, using principal component analysis and discriminant function analysis on hyperspectral images (Bonifazi et al., 2018, 2019). The prediction of PCB sources and factors influencing air concentration in contaminated buildings was modeled using field sampling data and principal component analysis (Kolarik et al., 2016). However, the implementation of neural network techniques (indicated in orange) in hazardous material management remains nascent, with limited examples in indoor radon and asbestos prediction. Convolutional neural networks have shown promise in identifying asbestos roofing from aerial imagery (Krówczyńska et al., 2020; Raczko et al., 2022), while artificial neural networks and recurrent neural networks were employed for estimating indoor radon exposure from meteorological and geological factors (Oni et al., 2022) and forecast and monitor the development of indoor radon concentrations (Khan et al., 2021; Valcarce et al., 2022) respectively.

Overall, these studies indicate that the data-driven applications identified have potential to support the processes outlined in the EU CDW Management Protocol,

particularly in material characterization, identification, and source separation. However, there is a noteworthy need of advanced predictive modeling for characterizing and identifying hazardous substances at the building level. The scope of existing studies is often confined to local contexts due to limitations in data accessibility and availability. Most research has focused on descriptive statistics of specific hazardous building components, utilizing various sources of environmental data, such as visual inspection guided by mobile app questionnaires (Govorko et al., 2017, 2018, 2019), asbestos product databases (Mecharnia et al., 2019), physical asbestos inventories (Krówczyńska et al., 2020; Raczko et al., 2022; Wilk et al., 2017, 2019), and demolition inspection reports (Franzblau et al., 2020). The exploration of pre-demolition audit inventories as input data for analyzing and modeling the presence patterns of hazardous substances remains uncharted in the literature. Given that the research silos focused on specific substances and the methodological constraints of descriptive analysis, environmental inventories offer a new avenue to address these challenges. They provide comprehensive inspection records of hazardous substances and materials across Swedish municipalities, presenting an opportunity to bridge existing gaps in the field.

3.1.2. Present Status and Opportunities in the Swedish CDW Sector

This section sums up findings from dialogues with industry via a workshop and an expert interview. The documentation from the workshop was systematically examined in a stakeholder need analysis, and the insights gleaned from expert interviews were compiled. To maintain clarity and focus, these results were organized thematically, centering on topics related to pre-demolition audits and decontamination practices. The section concludes with a comprehensive summary, synthesizing key findings from both the workshop and expert interviews.

Information availability of hazardous materials

The accessibility of information regarding hazardous materials is a crucial aspect of planning for CDW management. This planning typically involves synthesizing information from various sources, including desk studies – such as historical environmental inventories, building documents, and drawings – and field surveys, which encompass material, reuse, recycling inventories, and destructive sampling. Historical inventories and building registers are particularly valuable as they can indicate potential contamination in buildings accumulated over years of usage, thereby guiding auditors in environmental investigations. Additionally, information about building renovations, including the nature and timeline of changes, is vital for comprehensive pre-demolition audits. However, a significant challenge arises from the lack of data on renovation years and extent. This gap is often due to many renovations not being recorded in building permit systems. For instance, minor

alterations such as changing floor mats in the 1960s were typically not documented, whereas significant modifications, such as updating ventilation systems, required mandatory reporting. Subsequently, the sorted material and environmental inventories form the basis for subsequent procurement processes. BIM systems have not yet been extensively utilized for CDW planning in existing dwellings, possibly due to the unavailability or inaccessibility of such comprehensive information for older buildings.

The extent of hazardous material problems

The prevalence of hazardous materials in renovation and demolition projects is of significant concern. Most of the participants responded that they sometimes (43%) or often (29%) encounter hazardous materials in renovation or demolition projects, while none reported never (0%). These responses vary, likely reflecting differences in building types and ages, as well as the experience of the participants. The unexpected discovery of hazardous materials such as asbestos and PCB during renovations or demolitions poses additional risks, including the potential spread of contaminants to soil and groundwater. Therefore, having a robust construction management organization capable of making ad-hoc decisions is vital.

The impact of hazardous materials on project timelines and costs is significant and varies according to the types and amounts of materials discovered. Participants generally agree that projects involving hazardous material decontamination take considerably longer, sometimes double the time and cost, compared to projects without such materials. In extreme cases, these issues can delay project commencement, leading to substantial financial penalties. Such delays not only affect the initial demolition or decontamination stages but also have cascading effects on subsequent contractor schedules, necessitating additional resources for renegotiation. This is particularly critical in projects with fixed deadlines, such as hospital reconstructions, where schedules for demolition, decontamination, and rebuilding are tightly interlinked.

Pre-demolition auditors can also play a vital role in assisting with project timing adjustments in light of potential decontamination interventions. Currently, the quality of pre-demolition audits and environmental inventories varies widely across regions. This variation is attributed to a shortage of competent auditors amidst high market demand, coupled with inadequate quality control by authorities and property owners. The legislative requirements for auditor qualifications differ from hazardous materials. From a procurement standpoint, accurately estimating the extent of hazardous materials for project planning is challenging without high-quality inventories. There is a pressing need for detailed information from these inventories, regarding the extent and location of hazardous components, to be communicated to contractors and clients. Furthermore, there is a need for industry-wide education, to be conducted regularly, to standardize pre-demolition audit practices and inventory documentation. This standardization could be based on the Swedish Construction Federation's checklist. Additionally, feedback from

contractors and clients, both during and after audits, is crucial for auditors to refine their practices.

Hazardous material risk assessment and remediation

In construction projects, approximately 15% of the budget is typically allocated as a buffer to manage the unforeseen discoveries and risks associated with hazardous materials. Standard checklists have been developed for scenarios where hazardous materials are suspected. These include immediate work cessation, evacuation, and event reporting. However, determining the rational allocation of extra buffer to mitigate unknown risks in procurement is challenging. The extent of these additional costs varies depending on whether the project is privately or publicly funded. In public procurement, budgets may double due to inflated pricing from project partners, whereas private projects often involve fewer actors and typically offer a budget price rather than a fixed price. The phase of construction also influences the cost implications; for example, if hazardous materials are discovered during foundation digging, the cost can be significantly higher.

Most companies adhere to the regulations set by the Swedish Working Environment Agency, which include reporting incidents, material sampling, follow-up actions, and planning. It is crucial for clients to engage experienced contractors who are proficient in decontamination routines and proactive in maintaining work environment safety. The responsibility extends beyond just the entities performing decontamination; it is important to impose requirements on all project partners involved in exposure risk. Currently, the understanding and management of hazardous material risks are more prevalent among decontamination companies than demolition firms. There is a pressing need for enhanced regulatory requirements for demolition companies, including mandatory occupational education and certification, supported by project leaders.

Inventory of hazardous wastes

Pre-demolition audits in construction projects are fundamentally driven by systems thinking and a methodical approach to hypothetical searching. Auditors focus on understanding the reasons for the existence of specific hazardous materials, as this knowledge directs them to potential additional occurrences. For instance, asbestos has historically been used in buildings for its waterproofing, fireproofing, and electrical insulating properties, while PCB was used for its chemical stability, fire resistance, and insulating capabilities. It is, therefore, essential for procurement documents to mandate destructive sampling in pre-demolition audits, given that many hazardous materials only become detectable once demolition is underway. There have been instances where environmental inventories overlooked hidden hazardous materials, necessitating comprehensive reviews during and after renovations. To address this, supplementary inventories or post-demolition inventories are advisable, facilitating discussions with those responsible for the

initial environmental inventory. Additionally, some contractors conduct material reuse or recycling inventories to supplement pre-demolition audit inventories.

Different methods are employed to detect and quantify hazardous materials: PCB sealants are measured using scanners to determine their presence in meters, while PCB capacitors are counted. Asbestos detection typically involves polarized light microscopy or scanning electron microscopy to identify fiber types, with positive identification requiring at least 0.1% asbestos content, which can range between 5-100%. There are indicators for possible asbestos presence in fireproofing materials, with sealants often appearing similar but varying in fireproofing components. Therefore, sampling remains the only reliable method for detecting these hazardous substances. Besides, more synergies have been observed in detecting various asbestos-containing materials than PCB-containing ones across building types or construction periods. This is likely due to asbestos being more prevalent than PCB. Key factors in identifying these materials include the construction year, building type, location, and room type within the building.

Urban development policies also significantly influence the presence of hazardous materials in existing building stocks. In Stockholm and Gothenburg, a larger share of the multifamily building stocks were constructed during the Million Homes Program, characterized by modular design and factory production. The sealants used in joints between these modules often contained asbestos to enhance fire and waterproofing properties, only detectable upon dismantling the moisture protection. Notably, many buildings from the 1960s and 1970s have undergone renovations where original hazardous materials were removed. Some school buildings from this period were demolished due to poor energy performance or moisture damages.

The risk of secondary contamination is higher with PCB compared to asbestos, leading to more stringent regulation of its contamination levels. Furthermore, the tolerance threshold for PCB contamination in buildings is higher than that for soil contamination. PCB, along with Pulmonary Arterial Hypertension (PAH), can leak and cause severe environmental consequences, entailing costly post-handling processes. It is crucial to consider these factors in the evaluation and management of hazardous materials. A thorough understanding of the characteristics and historical usage of materials such as asbestos and PCB is vital in guiding their detection. Commonly found materials include asbestos-containing pipe insulation, cement panels, tile or clinker joints, floor mats, floor mat glue, and ventilation channels, as well as PCB-containing joints in multifamily houses or school buildings. PCB can also penetrate adjacent materials, leading to secondary contamination that may not be classified as hazardous but can impact occupant health. For PCB waste handling, disposal of PCB-contaminated gypsum board in landfills is prohibited. The characterization of asbestos and PCB materials in Sweden is listed below.

Asbestos-containing materials:

- **Pipe insulation:** Typically covered by plastics.
- **Valves:** Characterized by a square hole for natural ventilation.
- **Door and window insulation:** Found in fire doors, window putty, opaque windows, and balcony partition windows.
- **Windowsill:** Resemble marble imitation, difficult to remove.
- **Cement panel:** For external use, they often have hexagon patterns and are UV-resistant; internal panels are yellowish, used for wind protection. Commonly found in walls of school buildings, multifamily houses, and elevator shafts.
- **Tile and clinker joint:** Identified by the smooth edges of ceramic tiles, which are slightly thicker. Asbestos was occasionally added as an additive at construction sites.
- **Floor mat glue:** Visually identifiable in black or yellow colors. Black glue contains asbestos and PAH, capable of penetrating building materials up to 2 cm over time.
- **Floor mat:** PVC and vinyl floor mat may be re-glued with asbestos-containing floor mat glue. They often have multiple layers, including a patterned PVC film on the surface.
- **Ventilation channel:** Difficult to remove; typically, buildings have only one type of ventilation system, so testing one is usually sufficient.
- **Joint or sealant:** Used for fireproofing, especially in technical applications such as the joint between window frames and walls. Not associated with PCB joints, as asbestos joints are not found on facades.

PCB-containing materials:

- **Joint or sealant:** Sometimes found in hatches. PCB sealants in ventilation systems are yellowish and require cutting for detection.
- **Double-glazed sealed window:** Rare and often replaced due to low performance.
- **Capacitor:** Found in older lamps or burners.
- **Acrylic flooring:** Technically viable for about 10 years, few remain, typically used as non-slip floors in kitchens.

Pre-demolition audits are intricate endeavors that demand a high level of craftsmanship, skill, and experience. The decision-making process for selecting sampling points and material types is deeply rooted in a blend of professional knowledge, practical experience, and adherence to legal requirements. Auditors often rely on results from previous audits to inform their decisions regarding new sampling points. Acquiring proficiency in understanding buildings and their material use, mastering inventory techniques, and navigating relevant legislation is

a time-intensive process. In practice, auditors must consider practical issues and resource availability beforehand, utilizing all senses during field investigations.

Specific attention is required during asbestos sampling to ensure contamination protection and adherence to procedural guidelines. This includes preparing adequate tools, ensuring an ample supply of sample bags for different substances to prevent cross-contamination, and safe storage of collected samples. Consequently, pre-demolition audits are both costly and time-consuming. The cost for sampling is estimated at approximately 500-1000 SEK for each asbestos sample, 500 SEK for each PCB sealant sample, and 1500 SEK for each PCB concrete sample. The time and cost to conduct an environmental inventory vary depending on the type of building:

- **Single-family house:** Simple and fast inventory, typically requiring a maximum of one day of fieldwork and 1-2 days for a short report. The cost ranges from 15-30k SEK, with 3-5 samples typically collected.
- **Multifamily house:** Requires 1-5 days of fieldwork, with more extensive data collection and a 2-5 day reporting period. The number of samples can range from 15-50, incurring costs between 75-200k SEK.
- **School building:** For facilities accommodating 500-1000 students, with features such as elevators, large kitchens, dining rooms, gymnastic buildings, and shower rooms, the audit can take 2-5 days of fieldwork. Reporting may require 3-5 days, with 25-75 samples collected, and costs ranging from 100-250k SEK.
- **Office building:** The size and complexity of office buildings can vary greatly, influencing the duration of the audit (1-10 days of fieldwork) and reporting (2-10 days). The number of samples may range from 15-100, with associated costs between 75-350k SEK.

Decontamination of hazardous wastes

The decontamination of buildings contaminated with asbestos and PCB is governed by the guidelines AFS 2006:1 and SFS 2007:19. The procedural details are outlined on the website ”sanerapcb.nu,” which specifies the requisite steps for effective decontamination. Two primary methods are employed: isolating and decontaminating each location individually or sealing off a larger area for collective treatment. A critical aspect of this process is the prevention of dust migration to uncontaminated areas. The procedure concludes with an exhaustive dust-cleaning phase. To verify the absence of asbestos or PCB, post-cleaning sampling is conducted.

Furthermore, the decontamination process involves specific financial implications with a starting fee and a post fee. For asbestos, the costs are determined by the weight of the materials, ranging from 1500 to 2500 SEK per ton. This procedure requires prior registration, at least one day before the commencement of

remediation. In contrast, PCB decontamination mandates engagement with local municipal authorities. Contractors must submit sampling results and a detailed decontamination plan at least three weeks in advance to obtain the necessary permissions. Ultimately, property owners bear the responsibility for submitting final compliance reports.

Prospects for pre-demolition audit practice

The need for standardized and cohesive requirements in Swedish pre-demolition audits is imperative due to distinct environmental legislations governing contaminated land and buildings. Currently, the disjointed legislative framework concerning CDW often leads to ambiguity in assigning accountability between authorities, consequently shifting the responsibility of hazardous material management to property owners. Moreover, the lack of mandatory certification for conducting pre-demolition audits results in minimal client demand for inventory quality in Sweden. While auditors are required to undergo training for asbestos inventory, the educational provision for PCB inventory is limited to only half a day. To elevate the standard of pre-demolition audits, the introduction of a personal certification process complemented by comprehensive occupational training is recommended.

The integration of a machine learning-based decision support tool could significantly aid property owners in risk assessment and informed investment decisions. This tool could provide a predictive overview of demolition costs, offering valuable foresight. Although such a tool would not replace the necessity of pre-demolition audits, it could serve as a crucial adjunct to enhance the comprehensiveness of inventories. For regulatory bodies such as the Swedish National Board of Housing, Building, and Planning, the application of this tool necessitates rigorous validation. Additionally, the development of a machine-readable digital data model for inventories holds potential benefits beyond compliance, such as facilitating material reuse and recycling. The next step would involve the creation of a digital inventory protocol and an Application Programming Interface (API) for AI services. Collaborative development of a digital inventory template by municipalities and industry stakeholders could ensure that the data collated serves broader applications.

Summary

Perspectives from industry representatives reveal that environmental inventories conducted by qualified auditors are essential for mitigating risks associated with hazardous materials, forming a critical foundation for both decontamination and waste disposal planning. The environmental investigation, guided by a hypothesis-driven approach, relies on information sourced from building documents, drawings, and historical inventories. Key parameters such as the construction year, building usage, renovation history, specific location, and type of rooms in the building serve as indicators for the presence of hazardous materials. It is noted that certain hazardous materials may co-occur, such as asbestos in floor mats and adhesives, while others do not exhibit such synergies, as seen with asbestos and PCB joints. A thorough understanding of material characteristics and historical usage patterns is instrumental in detecting these materials.

Particularly, buildings erected during the Million Homes Programme are identified as having a higher likelihood of containing hazardous materials, especially in residential and school buildings within urban areas. Presently, approximately 15% of additional budgeting is allocated for unanticipated decontamination and disposal of hazardous materials. This figure is comparatively lower than the actual costs incurred in asbestos abatement, which account for 20.1% of the total demolition costs in residential dwellings in the City of Michigan (Franzblau et al., 2020). For clients and contractors, project delays often have more severe implications than extra costs, including penalties for contract breaches, postponement of subcontractor work, and equipment rental expenses. Given that the detection of several hazardous materials necessitates destructive sampling, it is recommended that the scope of the pre-demolition audit be contractually established to facilitate thorough decontamination planning and support safe hazardous material management through supplementary and post-inventory processes.

Moreover, the establishment of an industry-wide training or certification system, in tandem with a unified legislative framework, is proposed to enhance proficiency in hazardous material management and address the evolving demands of the CDW market. Additionally, the incorporation of data-driven applications, such as digital inventory protocols, machine learning techniques, and associated data models, can significantly contribute to the transition towards circular construction. These technological advancements have the potential to revolutionize various aspects of the industry, including material reuse and recycling, and the risk assessment of hazardous materials.

3.2. Data Preprocessing and Analysis (RQ 1)

This section examines the empirical documentation of existing pre-demolition audit inventories, with an emphasis on data transformation and systematization. In Conf I, a case study was featured that delved into the search for available building-related environmental information within inspection records. Following this, Paper II restructured and represented inventory information, then evaluating data usability through a specifically designed data assessment matrix. This data-driven approach not only enhances the understanding of the current state of pre-demolition audit inventories but also contributes to the development of more efficient methods for data handling and analysis in this domain.

3.2.1. Hazardous Material Detection Records in Inventories

The descriptive analysis of the pre-demolition audit inventories revealed disparities in both the availability and reliability of information across different types of inventories. The detail level of these inventories ranged from the most comprehensive in consultancy reports to the least in demolition plans, with protocols and control plans falling in between. Consultancy reports and protocols typically encompass exhaustive records of hazardous waste detection, adhering to the standards outlined in the "Industrial Resource and Waste Guidelines for Construction and Demolition" (Byggföretagen, 2019). These reports and protocols often detail the inventory of hazardous materials by type and quantity, particularly for larger or more complex buildings, through a combination of visual inspection and laboratory analysis of material samples collected during field surveys. The reports specify the scope of inspection, inventory extent, and information about the auditors involved.

In contrast, control plans and demolition plans tend to provide only general inspection records for simpler buildings, making it challenging to ascertain the completeness of inspection due to their aggregated or often missing information in free-text formats. The majority of reports focused on schools, commercial-, industrial-, office buildings, and multifamily houses, while single-family houses were predominantly covered in protocols, demolition plans, and control plans. A further analysis of auditor experience levels revealed that reports composed primarily by hazardous waste experts (96%) exhibited the highest data quality. Other inventories were predominantly conducted by contractors (55%), private individuals (27%), and hazardous waste experts (18%).

The analysis also highlighted discrepancies between inventories when evaluating the detection rates and missing values for hazardous substances. PCB, for instance, was detected in 63% of reports, 51% in demolition plans, 49% in protocols, and 19% in control plans. Conversely, asbestos detection varied

significantly across different document types, with the highest rates in reports (84%) and demolition plans (70%) in the Gothenburg renovated or demolished building stock, compared to lower rates in protocols (51%) and control plans (47%). A more granular investigation by components showed higher positive detection rates for nearly all asbestos and PCB materials. These findings suggest that relying on a single source for determining the presence of hazardous substances could introduce statistical uncertainties, as inventory types correlate with building classes. This association was further validated by recompiling detection rates according to building classes, revealing that the presence of hazardous materials is specific to building classes and should be considered when partitioning data for subsequent modeling.

Figure 3.3 presents the variation in the quantities of inventoried PCB and asbestos materials across different municipalities and building classes. PCB joints were the most commonly inventoried among other PCB materials, followed by double-glazed sealed windows, capacitors, and acrylic flooring across all building classes. As for asbestos, pipe insulation and tile or clinker were most frequently inspected, with asbestos valves being least common, consistent across regional building stocks. The majority of the hazardous material database inventories derived from multifamily houses, followed by single-family houses and non-residential buildings. Inspections of asbestos in floor mats, ventilation, and joints were less frequent in industrial buildings compared to other building classes. In particular, most inspected buildings were located in the Stockholm and Gothenburg municipalities, outnumbering those in Malmö and Kiruna by a significant margin due to different sizes of cities. These summative findings offer crucial insights into the data structure concerning regional and building compositions and were taken into account for later predictive result inference.

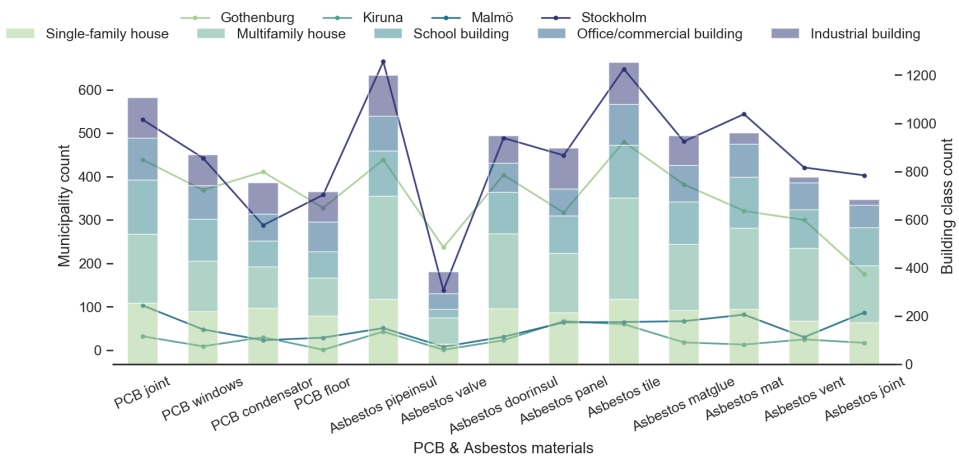


Figure 3.3. Detection records of PCB and asbestos materials shown by municipalities and building classes (data sourced from Paper IV).

The production of PCB and asbestos were common from the year 1930 to 1990, with their peak usage in construction occurring between 1965 and 1974. Figure 3.4 provides an insightful analysis of the correlation between the detection of asbestos and PCB substances across various construction years, underscoring the building stock most susceptible to contamination. The findings reveal a non-linear probability distribution, indicating that buildings constructed more recently, particularly after 1973, are less likely to contain either PCB or asbestos. Additionally, the data suggests an uneven distribution, with buildings erected between 1940 and 1955 likely to contain either PCB or asbestos. The likelihood of dual contamination from both asbestos and PCB becomes more pronounced in buildings constructed between 1955 and 1973. Following this period, the probability of detecting asbestos and PCB in buildings drops significantly and continues to decline towards the end of the 1980s. This analysis provides insights into the temporal trends of hazardous material usage in construction, informing targeted approaches for contamination assessment and remediation.

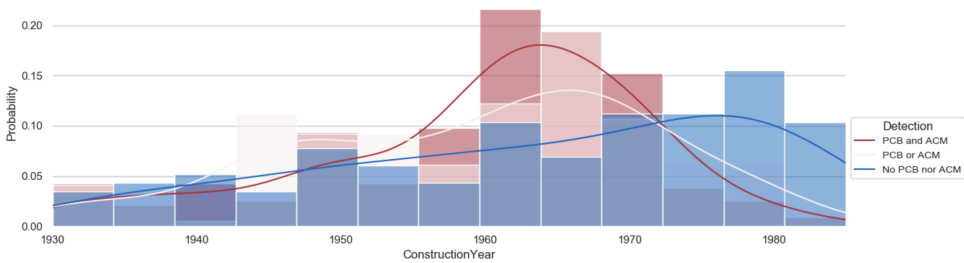


Figure 3.4. Probability distribution of buildings with single, double, and non-detection of PCB and asbestos materials across the construction year (data sourced from Paper IV).

Figure 3.5 presents stacked histograms to depict the frequency of detection for various asbestos and PCB materials across different buildings. This graphical representation offers insights into the prevalence of these hazardous substances in the building stock. The analysis of PCB materials reveals that 67.9% of the observed buildings contained no PCB materials. However, 23.7% of the buildings had one type of PCB material, while 8.4% contained at least two different PCB materials. Buildings constructed between 1955 and 1960 exhibited a significantly higher likelihood of containing multiple types of PCB materials compared to those built in the periods 1950-1980 and 1930-1950.

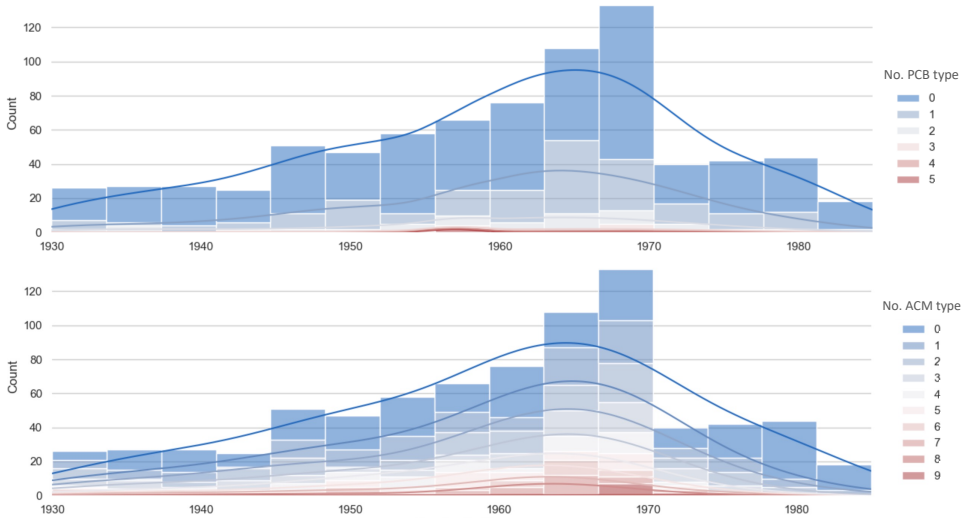


Figure 3.5. Stacked histograms of sample distributions for multiple detections of PCB (top two) and asbestos materials (bottom two) across the construction year (data sourced from Paper IV).

In contrast, the observations for asbestos materials showed a different pattern. Only 35.3% of buildings had no asbestos materials. Around 17.8% contained at least one type of asbestos, 14.8% had two types, and 10.7% contained three types of asbestos materials. Remarkably, approximately 21.4% of the observations indicated the presence of at least four different asbestos materials. A trend of increasing asbestos contamination was particularly evident in buildings constructed between 1955 and 1975, with the peak probability of detecting multiple types of asbestos occurring between 1965 and 1975. These distinctive patterns in the probability distributions underscore the construction year as a critical factor in determining the presence of hazardous building materials. The temporal analysis reveals specific periods with elevated likelihood of contamination, providing valuable insights for targeted hazardous material management in buildings from these eras.

The hierarchical cluster matrix illustrated in Figure 3.6 provides a comprehensive analysis of the associations between the presence of PCB and asbestos-containing materials in various buildings. The results predominantly indicate low correlations both within and between different types of PCB and asbestos components. However, exceptions were observed where significant statistical correlations were identified. A noteworthy correlation was found between asbestos ventilation channels and asbestos tiles or clinker, exhibiting a coefficient of 0.62. Similarly, a correlation coefficient of 0.61 was observed between PCB joints and asbestos tiles or clinker. These correlations suggest a concurrent presence of these materials, particularly in specific types of buildings, presumably multifamily houses. Such synergies in material usage patterns, evident from the

analysis, could be attributed to the construction practices prevalent during certain periods. These findings are instrumental in enhancing the predictive inference of hazardous materials in the Swedish building stock. By understanding the distinct presence patterns of hazardous materials, especially in specific building classes, this analysis facilitates more targeted and effective strategies for hazardous material detection and management in the renovation and demolition processes.

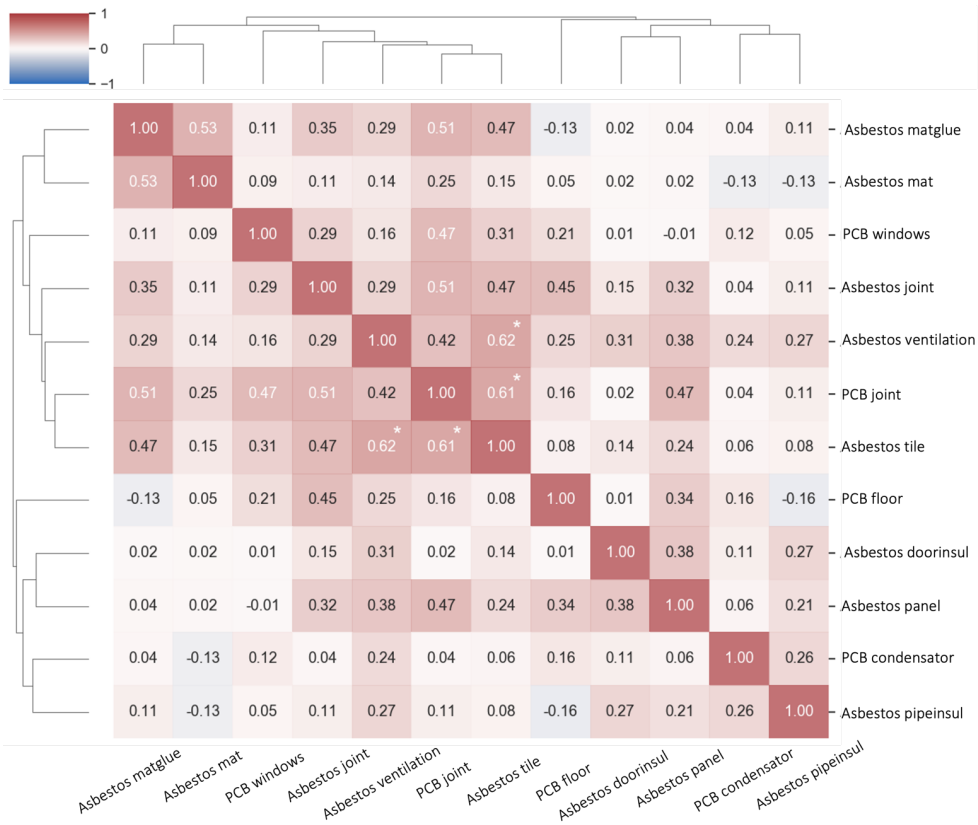


Figure 3.6. Variable correlation between asbestos and PCB materials. Coefficients with statistical significance are marked with asterisks (data sourced from Paper IV).

3.2.2. Evaluation of Inventory Records

Evaluating the representativeness and quality of data from inventory records is a crucial step in predictive modeling, particularly for determining the suitability for using them as training data. Paper II revealed that the inventory data encompassed approximately 2.2% of Gothenburg's building stock constructed between 1929 and

1982, as per building count. Essential building parameters including construction year, renovation year, floor area, and number of floors were extracted from both the training and prediction sets for comparative analysis of their distributions. Analysis using normalized density distributions and box plots showed that the median construction and renovation year intervals (1962-1964 and 1996-1998, respectively) were similar across both data subsets. It was observed that most buildings in the prediction set were constructed later, predominantly in the 1970s, with significant renovations around 1990 and 2005. The training set exhibited broader ranges in floor area and number of floors, likely due to the extensive demolition and extension activities in low-rise and simple buildings, such as storages and garages.

In Paper IV, the research extended its scope to encompass four distinct municipalities, resulting in a considerably improved alignment of building characteristics across the various data subsets for enhanced validity of the study's findings. This expansion allowed for a more comprehensive and representative analysis of building attributes, such as construction year, number of floors, basements, and stairwells. However, the training set still showed an overrepresentation of larger buildings with more apartments, indicating a potential bias in the data. This was further confirmed by differences in building class distributions between subsets: the training set had a nearly equal distribution among various building classes (18% single-family houses, 19% multifamily houses, 24% school buildings, 20% office and commercial buildings, and 19% industrial buildings), whereas the prediction set predominantly comprised 80% of single-family houses and 16% of multifamily houses, with a minimal representation of non-residential buildings (4%).

Subsequent analysis focused on identifying building subgroups within the training set that exhibited high quality and quantity of data, assessing their suitability for predictive modeling. The development of a data assessment formula and matrix, as introduced in Paper II, facilitated the data evaluation of hazardous materials across different building classes based on inventory types and quantities. This approach was applied in Papers II-IV. Top assessment scores for PCB and asbestos-containing materials were consistently found in school buildings, commercial buildings, multifamily houses, and industrial buildings. However, the ranking of hazardous materials varied depending on the characteristics of inventoried buildings across different sampled municipalities (Gothenburg in Paper II, Gothenburg and Stockholm in Paper III, and an expanded set including Gothenburg, Stockholm, Malmö, and Kiruna in Paper IV). Despite variations in the training set, certain materials such as asbestos pipe insulation, asbestos tiles or clinker, asbestos door or window insulation, asbestos ventilation channels, and PCB joints consistently achieved high scores, supported by a substantial number of inventories from reports and protocols.

3.3. Prediction of Hazardous Materials (RQ2 & 3)

This section synthesizes the outcomes of predicting asbestos and PCB-containing materials in buildings constructed from 1930 to 1985, elucidating the machine learning pipelines developed specifically for this research context. Paper III details preliminary efforts to forecast the presence of a single type of both asbestos and PCB material. Conversely, Paper IV introduces a comprehensive machine learning framework designed for predicting various types of asbestos and PCB materials across distinct building categories. The lead models were then applied to data from buildings yet to be investigated, enabling an estimation of the likelihood and spatial distribution of residual hazardous materials within the Swedish building stock.

3.3.1. Predictive Modeling for Hazardous Building Materials

The aggregated results from the application of machine learning and neural network models, as detailed in Table 3.1, demonstrate the ability to discern presence patterns for a range of PCB and asbestos-containing materials. Values highlighted in bold represent the superior performance of lead models for specific hazardous materials. Key data preprocessing strategies, such as data resampling, augmentation of minority classes, and adjustment of sample weights, proved effective in mitigating label imbalance during model training. The models exhibited consistent performance metrics, particularly in terms of Area Under the Curve (AUC) and F1 scores, while maintaining close error rates between training and validation sets, lending greater confidence to their performance evaluation.

Among the various models tested, tree ensemble models – including Distributed Random Forest or Extremely Randomized Trees (DRF/XRT), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Stacked Ensemble models – generally outperformed others. High AUC scores were achieved for predicting the presence of asbestos door and window insulation (0.93), asbestos ventilation channels (0.90), PCB joints (0.88), asbestos floor mat glues (0.88), and PCB capacitors (0.86) in residential buildings. On the contrary, the predictive performance in non-residential buildings was marginally lower, with models for asbestos door insulation, PCB capacitors, asbestos joints or sealants, PCB joints, and asbestos pipe insulation yielding AUC scores of 0.85, 0.76, 0.76, 0.75, and 0.75, respectively. The overlap in the detection of these hazardous materials across different building types underscores the prevalence of their distinct presence patterns within the Swedish building stock.

Table 3.1. Performance evaluation of predictive models for PCB and asbestos materials prediction using the AUC metric (1e-2), as detailed in Papers III-IV.

Algorithm*	L	S	K	C	G	R	B	X	N	E
All buildings										
PCB joint	-	-	-	-	64	76	**78**	74	66	-
PCB double-glazed window	-	-	-	-	-	-	-	**60**	-	60
PCB capacitors	-	-	-	-	73	**82**	**82**	79	80	-
PCB acrylic floor	-	-	-	-	24	41	61	48	44	**66**
ACM pipe insulation	-	-	-	-	65	79	**80**	78	73	-
ACM door insulation	-	-	-	-	74	85	**84**	83	80	-
ACM panel	-	-	-	-	68	**71**	70	66	**71**	-
ACM tile/clinker	-	-	-	-	57	73	**75**	73	65	-
ACM floor mat glue	-	-	-	-	64	73	**75**	71	68	-
ACM floor mat	-	-	-	-	-	-	-	**63**	-	62
ACM ventilation channel	-	-	-	-	-	79	78	76	75	**80**
ACM joint/sealant	-	-	-	-	70	77	**80**	75	69	-
Residential buildings (single-family and multifamily houses)										
PCB joint	-	-	-	-	84	81	**88**	81	-	85
PCB double-glazed window	-	-	-	-	56	-	-	**61**	-	46
PCB capacitors	-	-	-	-	**86**	84	82	66	-	79
PCB acrylic floor	-	-	-	-	-	-	-	-	-	-
ACM pipe insulation	62	79	82	96	72	**81**	74	75	80	78
ACM door insulation	for multifamily house				73	88	**93**	84	70	90
ACM panel	-	-	-	-	76	**84**	77	75	65	82
ACM tile/clinker	-	-	-	-	55	74	**83**	82	51	80
ACM floor mat glue	-	-	-	-	69	-	-	**88**	68	**88**
ACM floor mat	-	-	-	-	46	-	-	**61**	44	54
ACM ventilation channel	-	-	-	-	84	89	85	83	74	**90**
ACM joint/sealant	-	-	-	-	-	**78**	79	74	-	75
Non-residential buildings (school, office/commercial, industrial buildings)										
PCB joint	86	83	78	99	58	-	-	**75**	-	72
PCB double-glazed window	for school building				**66**	68	60	66	-	66
PCB capacitors	-	-	-	-	75	-	-	**76**	-	75
PCB acrylic floor	-	-	-	-	41	-	-	**68**	-	37
ACM pipe insulation	-	-	-	-	65	76	**75**	73	-	74
ACM door insulation	-	-	-	-	80	-	-	**85**	-	85
ACM panel	-	-	-	-	57	**65**	62	59	-	65
ACM tile/clinker	-	-	-	-	60	67	**69**	63	-	66
ACM floor mat glue	-	-	-	-	70	72	**73**	71	-	69
ACM floor mat	-	-	-	-	54	67	**68**	65	-	67
ACM ventilation channel	-	-	-	-	70	**74**	68	66	-	72
ACM joint/sealant	-	-	-	-	61	72	**76**	73	-	73

*Abbrevidation of algorithms: L (logistic regression, LR), S (support vector machine, SVM), K (*k* nearest neighbors, *k*-NN), C (categorical boosting, CatBoost), G (generalized linear model, GLM), R (distributed random forest or extremely randomized trees, DRF/XRT), B (gradient boosting, BG), X (extreme gradient boosting, XGboost), N (deep neural network, DNN), and E (stacked ensemble).

The performance variances observed in the prediction models, as detailed in Papers III and IV, can be attributed to the expanded training data size and feature sets. In Paper III, the categorical boosting models demonstrated exceptionally high performance compared to other algorithms, suggesting a potential overfitting issue. On the other hand, neural networks, due to their less-than-ideal performance and similarity to simpler models such as generalized linear models that were deemed less suitable for training with structurally small datasets. The feature sets used in these studies also varied. Paper III utilized supplementary categorical variables such as the city and EPC category as training features. In contrast, in Paper IV, these attributes were replaced with the postcode, building category, and building types, in addition to derived features such as building physical footprint, area per stairwell, and area per apartment. The learning curve analysis indicated that model performance improved with increasing data size, achieving optimal results with a minimum of 100 observations. Paper IV findings suggested that simple models with approximately 20-55 trees and 5-13 depths were optimal, balancing the risks of overfitting and underfitting.

Furthermore, the aggregated feature importance based on lead models, as illustrated in Figure 3.7, highlighted that floor area, building physical footprint, construction year, postcode, and the number of floors collectively contributed to predicting PCB and asbestos materials in residential buildings. These parameters were instrumental in characterizing building categories and typologies, with the postcode particularly relevant to variations in building stocks across different geographical regions. For non-residential buildings, the construction year and postcode emerged as the most critical features, exerting significant impact magnitudes relative to others. Interestingly, the scaled feature importance varied between hazardous materials and building categories. Some materials had a dominant predictor, while others were influenced by an aggregation of features. Moreover, even the same materials exhibited different predictor sets based on the building category. Given these insights, it is advisable to develop individual predictive models tailored to specific regions and building categories, provided there is a sufficient volume of data for such data partition.

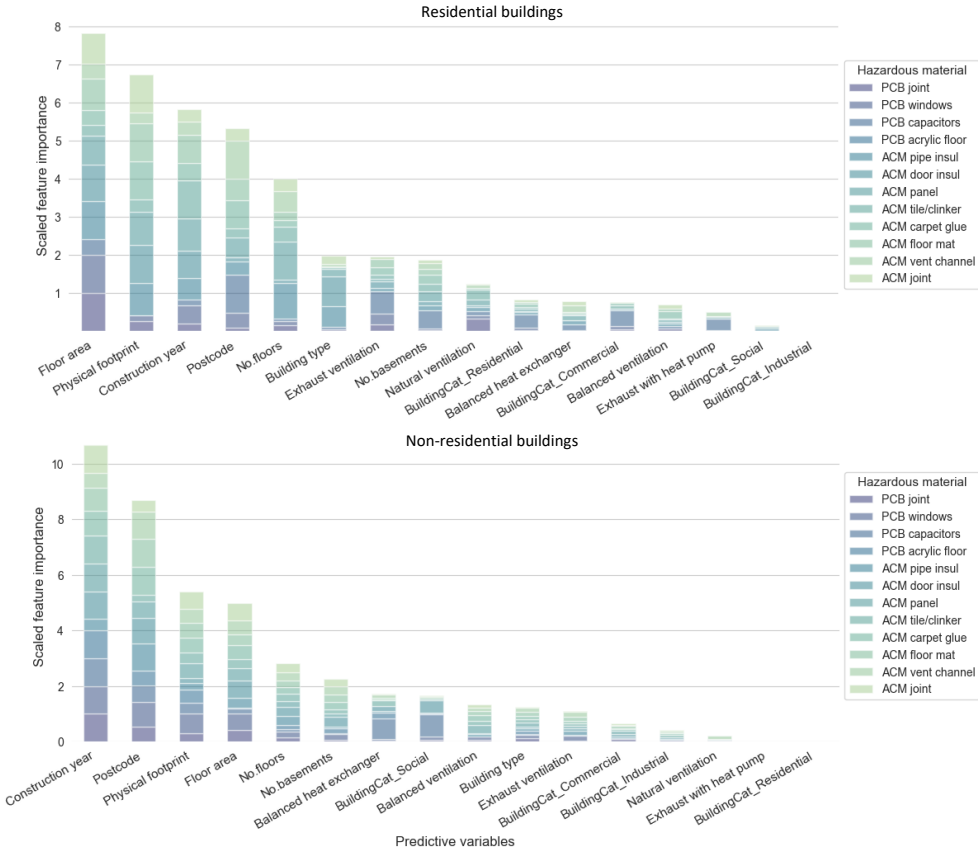


Figure 3.7. Aggregated feature importance of leader models for PCB and asbestos material prediction in residential and non-residential buildings (data sourced from Paper IV).

Afterwards, SHAP values were computed, aiding in the interpretation of the predicted presence of asbestos and PCB materials in both residential and non-residential buildings, as depicted in Figure 3.8. The analysis revealed significant feature impacts of leader models for specific hazardous materials in residential and non-residential buildings, illustrated in Appendix II. In residential buildings, significant impacts were observed for PCB joints, double-glazed sealed windows, asbestos in door and window insulation, tiles or clinker, floor mat glue, floor mats, and ventilation channels. In non-residential buildings, most asbestos materials, excluding ventilation channels, demonstrated pronounced feature impacts. Other materials exhibited less significant distinctions. The detailed patterns of their presence are identified as follows:

- **PCB joints**: Identified predominantly in medium to large post-war residential buildings with medium building footprints but lacking natural ventilation, also

commonly detected in post-war non-residential buildings in urban areas with medium to large footprints and floor areas, and balanced ventilation systems.

- **PCB double-glazed sealed window**: Found in large residential buildings constructed post-1970s in rural areas with basements, but without natural or exhaust ventilation systems, also present in smaller multi-storey post-war non-residential buildings in urban settings with medium footprints, basements.
- **PCB capacitor**: Detected in medium to large residential buildings with exhaust ventilation and medium footprints. Early-built commercial or industrial buildings without balanced ventilation also showed occurrences.
- **Asbestos pipe insulation**: Common in medium to large post-war multifamily houses with multiple stories and significant footprints, also observed in medium to large multistory non-residential buildings from the early period, particularly in urban areas with exhaust or balanced ventilation.
- **Asbestos door and windows insulation**: Predominantly found in medium to large post-war multifamily houses with large footprints and exhaust ventilation, also seen in medium to large multi-story post-war commercial or industrial buildings in urban areas with basements but without exhaust ventilation.
- **Asbestos panel**: Detected in multi-story residential buildings built later with medium to large footprints, also found in post-war non-residential buildings in urban areas with balanced ventilation and significant footprints and floor areas.
- **Asbestos tile and clinker**: Identified in multi-story residential buildings with natural ventilation and substantial footprints, also found in post-war non-residential buildings, particularly industrial buildings, with large footprints lacking balanced ventilation with heat exchangers, but present in urban areas with balanced ventilation.
- **Asbestos floor mat glue**: Appeared in medium early-built residential buildings in urban areas with medium to large footprints and exhaust ventilation, but not balanced ventilation, also presented in large post-war non-residential buildings in urban areas with basements.
- **Asbestos floor mat**: Found in small, low-rise residential buildings in urban areas built post-1950s, with medium footprints and basements, but without exhaust ventilation, also observed in large multi-story post-war industrial buildings in urban areas with significant footprints and basements, but lacking exhaust ventilation.
- **Asbestos ventilation channel**: Common in small, low-rise post-war residential buildings with balanced ventilation with heat exchangers or exhaust ventilation with heat pumps, also presented in medium-sized post-war non-residential buildings.
- **Asbestos joints**: Detected in medium to large residential buildings with significant footprints, also found in multi-story post-war non-residential buildings with large footprints, basements, and balanced ventilation.

3.3.2. Hazardous Material Prediction in the Building Stock

An examination of the sample distribution was conducted to assess the representativeness of the inventoried buildings in relation to the regional building stock with valid postcodes, prior to initiating prediction. The 287 sampled residential buildings constituted approximately 0.21% of the total residential building stock, which numbers 95,344. Despite the relatively small sample size, the proportional representation of municipalities within the sample was fairly consistent (40% Stockholm, 40% Gothenburg, 18% Malmö, 2% Kiruna in the training set; 41% Stockholm, 41% Gothenburg, 17% Malmö, 1% Kiruna in the prediction set). In contrast, about 1% of non-residential buildings, totaling 3,788, were sampled. The sampling exhibited a significant oversampling of Stockholm's non-residential buildings in both the training set (51% Stockholm, 32% Gothenburg, 15% Malmö, 2% Kiruna) and the prediction set (41% Stockholm, 35% Gothenburg, 23% Malmö, 1% Kiruna)

Figure 3.8 presents a comparison of the normalized density distributions for the training and prediction sets across various building categories, considering key features such as construction year, floor area, and building physical footprint. The analysis indicated that the training set tended to oversample residential buildings constructed between 1945 and 1970, and non-residential buildings built before 1970. Conversely, it undersampled smaller residential buildings with floor areas and physical footprints less than 1000 m², as well as non-residential buildings with floor areas below 4000 m². However, the distribution of building physical footprints in the training set aligns closely with that in the prediction set for non-residential buildings. This alignment suggests that the buildings inventoried in the four municipalities predominantly predate the bans on PCB (1973-1978) and asbestos (1975-1982). Consequently, this skew in the sampling may result in higher positive detection rates and frequencies of detecting multiple hazardous materials in the statistics than their actual prevalence rates in the overall Swedish building stock.

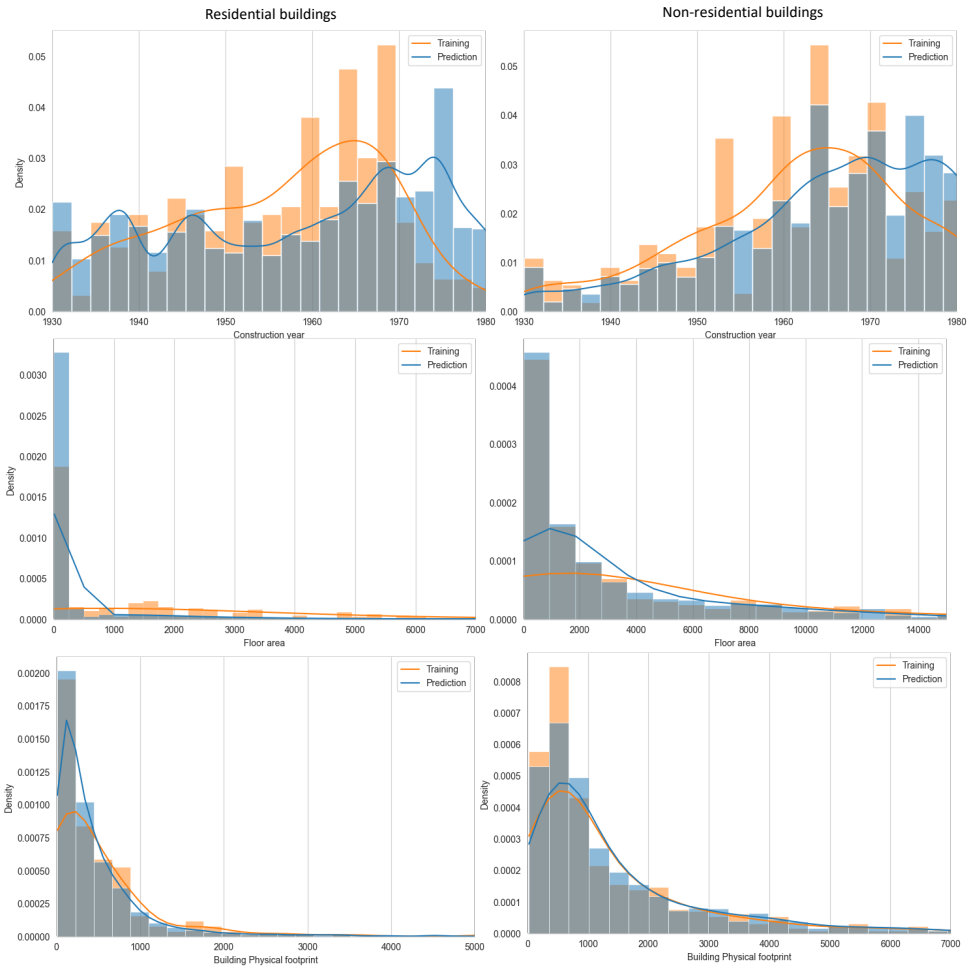


Figure 3.8. Normalized density distribution of training and prediction sets per building category by construction year, floor area, and building physical footprint (data sourced from Paper IV).

In Paper IV, the estimation of the presence of specific asbestos and PCB materials within regional building stocks was conducted at two different scales: global (encompassing the entire dataset) and local (focusing on individual cases). At the global scale, the prediction models generated binary labels, which were used to compute aggregated statistics for each municipality and building category. Table 3.2 provides a comparative analysis of the positive detection rates of inventoried buildings in the training set against the estimated rates for non-inventoried buildings in the prediction set. Additionally, the table includes details on the performance of each model and the data size for each subgroup, aiding in the assessment of uncertainty in the prediction outcomes.

The findings indicated that the predicted proportions of buildings contaminated with hazardous materials were generally lower than those derived from statistical analysis, particularly in the case of residential buildings. This observation aligns with the dense concentration of post-war buildings noted in Figure 3.9. When comparing the detection rates of hazardous materials across all building categories, the rates in the prediction set were lower and more closely aligned with those found in residential buildings. This is due to the fact that residential buildings account for nearly 95% of the overall building stock, whereas they represent only around 37% of the observations in the training set. The discrepancy in these proportions suggests that a higher number of non-residential buildings, compared to residential buildings, have undergone renovation or demolition with detailed inventories in the past decade in the municipalities included in the study. This could also imply a more comprehensive documentation of hazardous materials in non-residential buildings, impacting the prediction outcomes and their interpretation.

Table 3.2. Comparison between the positive detection rates of asbestos and PCB materials between the statistics of inventoried buildings (I) and prediction of non-inventoried buildings (P), as detailed in Paper IV.

Building	Stockholm		Gothenburg		Malmö		Kiruna		Total	
	I	P	I	P	I	P	I	P	I	P
Residential (N)	114	39,248	115	39,225	6	16,725	52	146	287	95,344
Non-residential	254	1,563	161	1,324	74	885	10	16	501	3,788
All (N)	368	40,811	276	40,549	80	17,610	62	162	788	99,132
PCB capacitor (%)										
Residential (AUC= 0.86)	0.33	0.10	0.27	0.04	0	0.01	1.00	0.78	0.36	0.02
Non-residential (AUC= 0.76)	0.53	0.30	0.62	0.58	0.18	0.48	0.80	0.56	0.56	0.44
All buildings (AUC= 0.82)	0.49	0.12	0.50	0.19	0.17	0.28	0.93	0.99	0.51	0.18
Asbestos pipe insulation (%)										
Residential (AUC= 0.81)	0.81	0.63	0.52	0.43	1.00	0.77	0.92	0.31	0.72	0.57
Non-residential (AUC= 0.75)	0.65	0.98	0.45	0.83	0.83	0.98	1.0	0.75	0.61	0.92
All buildings (AUC= 0.80)	0.71	0.56	0.48	0.28	0.84	0.56	0.94	0.81	0.65	0.44
Asbestos door and window insulation (%)										
Residential (AUC= 0.91)	0.81	0.20	0.43	0.10	NA	0.12	1.00	0.12	0.67	0.14
Non-residential (AUC= 0.85)	0.60	0.57	0.47	0.51	0.87	0.80	0.67	0.25	0.58	0.60
All buildings (AUC= 0.84)	0.67	0.23	0.46	0.14	0.87	0.23	0.90	0.21	0.61	0.19

At the local scale, the prediction models were employed to ascertain the likelihood of hazardous material presence in individual buildings, with results visualized both on aggregated and individual building bases. Figure 3.9 displays the estimated probability distributions for the presence of PCB capacitors, asbestos pipe insulation, and door and window insulation, categorized by building types and plotted against the construction year. These distributions are annotated with mean values and 95% confidence intervals to provide a clearer statistical perspective. The analysis revealed that non-residential buildings generally exhibited a higher probability for hazardous material contamination compared to residential buildings across all three evaluated cases. Residential buildings showed a comparatively greater likelihood of containing asbestos door and window insulation than asbestos pipe insulation or PCB capacitors. However, a noticeable downward trend in the predicted probability of containing these hazardous materials was observed post-1970s. Interestingly, older buildings constructed between 1930 and 1945 did not exhibit significantly lower probabilities than those built during the post-war period, a pattern potentially attributes to renovations undertaken in the 1970s.

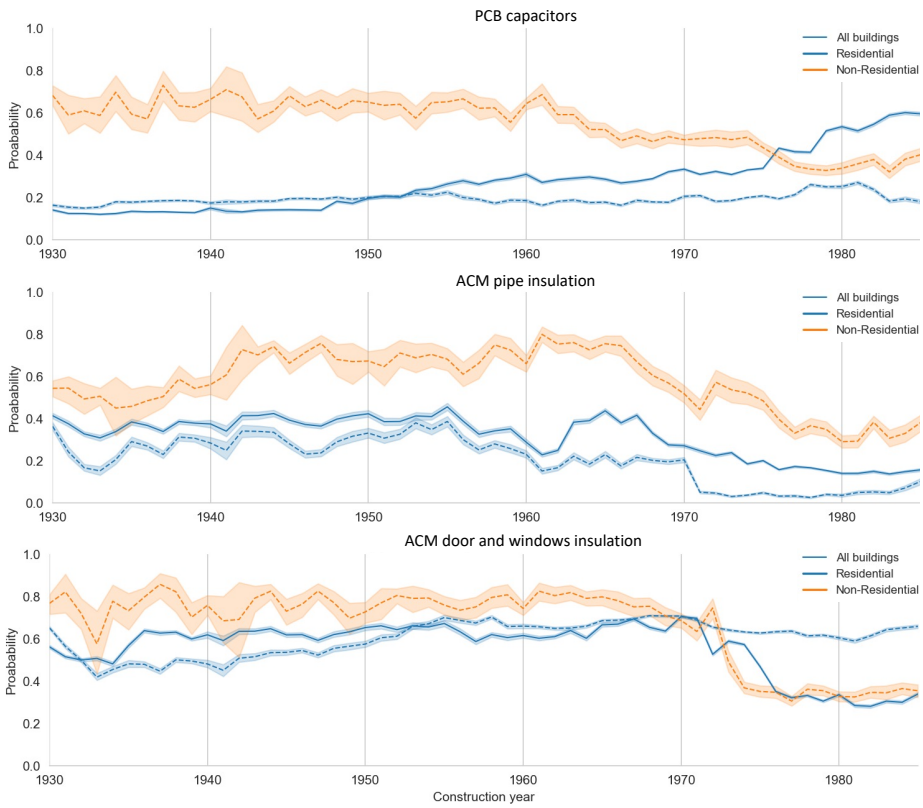


Figure 3.9. Predicted probability distribution of selected asbestos and PCB materials along the construction year by all buildings (in continuous line) and building categories (in dashed line), as detailed in Paper IV.

Furthermore, the research extended to the creation of hazardous material atlases for Stockholm's public housing, utilizing building footprint maps. These atlases facilitated the mapping and geospatial analysis of hazardous material probabilities at the district level. Through a color-coded building schema, potential correlations between the presence of hazardous materials and clusters of buildings more prone to contamination were discerned. This approach provides a novel perspective on understanding and visualizing the geospatial distribution of hazardous material probabilities within urban environments.

3.4. Estimation of Radioactive Substances (RQ2 & 3)

This section delves into the feasibility of transferring the proposed data-driven prediction methodology, originally developed for hazardous materials, to the prediction of radioactive substances in buildings. Specifically, Paper V investigates the potential of predicting the presence of radioactive concrete, employing an alternative statistical approach based on Bayesian network learning. Furthermore, in Conf II and Paper VI, an exploration by applying machine learning regression and multi-class classification techniques, respectively, is extended. These methods are utilized to forecast indoor radon concentrations and their distribution intervals within the national building stock.

3.4.1. Predictive Modeling for Radioactive Substances

The investigation focused on the prevalence and impact of radioactive concrete-containing materials in the existing building stock constructed between 1930 and 1980. It is well-known and also proved that buildings with radioactive concrete exhibited higher indoor radon concentrations, as demonstrated on a logarithmic scale for a broader range analysis, detailed in Figure 3.10. The average indoor radon concentration in buildings devoid of radioactive concrete was significantly lower, recorded at 94 ± 6 Bq/m³, compared to 319 ± 31 Bq/m³ in buildings with radioactive concrete. This difference was more pronounced in single-family houses (286 Bq/m³), followed by multifamily houses (99 Bq/m³), school buildings (45 Bq/m³), and other non-residential buildings (125 Bq/m³), as depicted in Figure 3.9. The detection rates of radioactive concrete varied by building type: approximately 14% in single-family houses, 49% in multifamily houses, 9% in school buildings, and 24% in other non-residential building categories. Records pertaining to the specific components containing radioactive concrete were often vague. However, where information was available, radioactive concrete was most frequently identified in walls, facades, floors or foundations, and other structural elements such as stairwells, chimneys, and air-conditioning service areas.

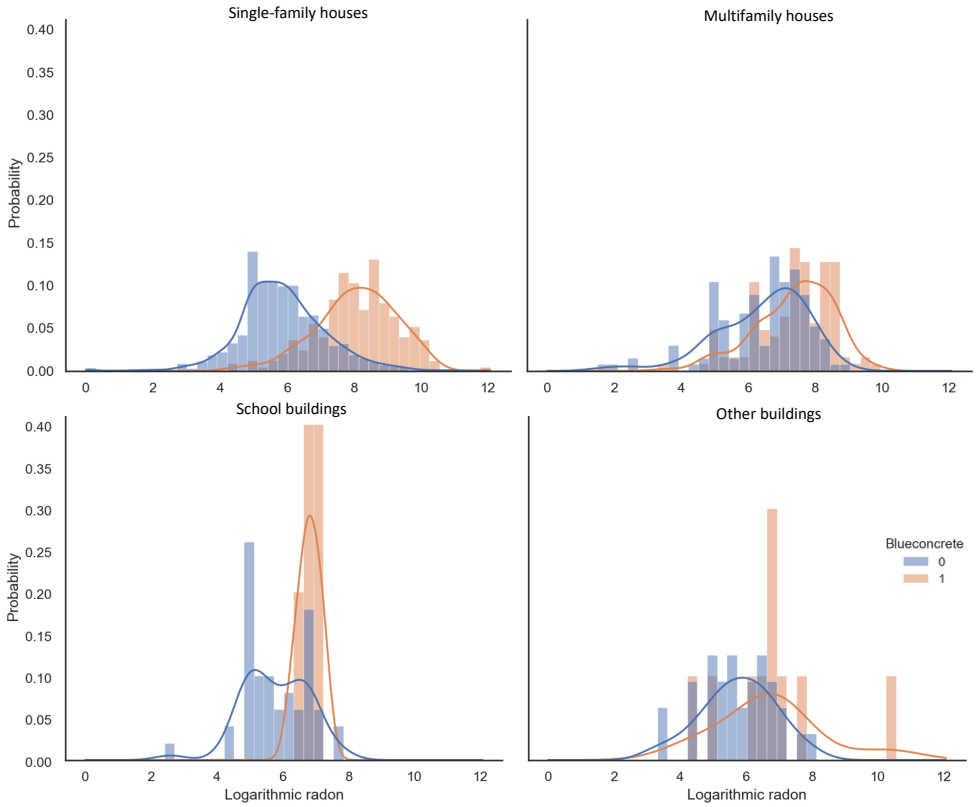


Figure 3.10. Probability distribution of logarithmic indoor radon concentration by building classes with and without radioactive concrete detection (data sourced from Paper V).

The correlation matrix presented in Paper V revealed a significantly negative correlation between the geographical proximity to radioactive concrete manufacturing plants and the presence of radioactive concrete in residential buildings. Specifically, the correlation coefficients were -0.89 for single-family houses and -0.87 for multifamily houses, indicating a higher prevalence of radioactive concrete in buildings located closer to these factories. In multifamily houses, the number of basements also exhibited a negative correlation (coefficient is -0.40). Conversely, positive correlations were observed with the construction year (coefficient is 0.37), number of apartments (coefficient is 0.26), and floor areas (coefficient is 0.21) in multifamily houses. Additionally, the number of floors in school buildings showed a strong positive correlation (coefficient is 0.71) with the presence of radioactive concrete. These findings were substantiated through diagnostic analysis of the Bayesian networks developed during structural learning. This analysis highlighted the average distance to radioactive concrete manufacturing plants as a pivotal attribute affecting the presence of radioactive

concrete across different building classes. This factor, in combination with building class, construction year, and floor area, was particularly influential. The directed acyclic graphs from structural learning further revealed the interconnections between construction year, number of basements, floor area, and building class.

In Table 3.3, a compilation of the conditional probability distributions derived from the parametric learning of models with the highest Bayesian Information Criterion (BIC) scores is shown. The compiled data indicate an average detection rate of radioactive concrete in 36% of the observations. Buildings most likely to contain radioactive concrete were those constructed between 1968 and 1975, especially multifamily and other non-residential buildings, excluding schools, with basements and located 300-600 km from historical radioactive concrete production sites. However, factors such as floor areas, number of floors, stairwells, and number of apartments exhibited no direct dependency on the presence of radioactive concrete. These findings are consistent with the patterns identified in previous exploratory data analyses performed in Paper V.

Table 3.3. Aggregated conditional probability distributions of Bayesian network models learned from radioactive concrete detection records, as detailed in Paper V.

Variable	Value representation	Radioactive concrete	
		Positive	**Negative**
Dataset		36%	64%
Construction year	1930-1955	23%	77%
	1955-1960	40%	60%
	1960-1968	60%	40%
	1968-1975	30%	70%
	1975-1980	12%	88%
Average distance (km)	below 300	62%	38%
	300-600	83%	17%
	above 600	2%	98%
Basements/ Building class	No/Single-family house	8%	47%
	Yes/Single-family house	14%	44%
	No/Multifamily house	58%	5%
	Yes/Multifamily house	53%	37%
	No/School building	16%	27%
	Yes/School building	6%	9%
	No/Other non-residential building	19%	21%
	Yes/ Other non-residential building	28%	11%

In addition to the factors contributing to indoor radon levels, their distribution was also quantified across different building classes. As reported in Paper VI, the mean annual indoor radon concentration in Swedish buildings was 110 ± 1 Bq/m³, with approximately 12.4% of buildings exceeding the 200 Bq/m³ reference level. Particularly, single-family houses (118 ± 2 Bq/m³) recorded higher than average indoor radon levels, with 13.9% surpassing the reference threshold, the highest proportion among all building classes. In contrast, school buildings exhibited the lowest mean indoor radon concentration. Non-residential dwellings frequently registered extremely high indoor radon levels (above 500 Bq/m³). These observations align with findings presented in Conf II, which utilized heat maps to demonstrate the distribution of indoor radon levels across various building categories. Buildings particularly vulnerable to high indoor radon concentrations were identified as those constructed between 1940 and 1960, with natural ventilation, and located in areas with elevated uranium content. The correlation study in Conf II and subsequent data analysis and visualization in Paper VI further detailed the relationships between indoor radon levels and a range of building parameters, geogenic factors, and geographic attributes for each building class.

Predictive modeling for indoor radon concentration estimation involved regression techniques. The performance of multivariate adaptive regression splines and random forest models, with r-squared values of 0.13 and 0.24 respectively, was in line with existing literature but not entirely satisfactory. Consequently, an alternative approach using multi-class classification was explored. This involved categorizing indoor radon levels into three intervals, in accordance with current legislative guidelines: low level (below 200 Bq/m³), medium level (200-400 Bq/m³), and high level (above 400 Bq/m³). By applying the SMOTE technique for label imbalance adjustment, the XGBoost models demonstrated significant improvement in prediction performance (F1 scores for single-family houses is 0.93, multifamily houses is 0.95, school buildings is 0.94, other non-residential buildings is 0.96), outperforming the deep neural network models (F1 scores for single-family houses is 0.66, multifamily houses is 0.74, school buildings is 0.64, other non-residential buildings is 0.68). The confusion matrix for the XGBoost models, presented in Table 3.4, indicated average error rates ranging from 4.3% to 6.5% across different building classes. These errors predominantly occurred in the misclassification of medium-level indoor radon concentrations as low level. The misclassification rates for high indoor radon levels were relatively low in residential buildings but increased in non-residential buildings.

Table 3.4. Confusion matrix of the XGBoost models for indoor radon prediction, as detailed in Paper VI.

Indoor radon level		True label (N)			
		Low	**Medium**	**High**	**Error rate**
Predicted label	**Single-family houses**				
	Low	820	45	18	7.1%
	Medium	79	788	16	10.8%
	High	8	5	871	1.5%
	Average	907	838	905	6.5%
	Multifamily houses				
	Low	3261	116	23	4.1%
	Medium	228	3100	62	8.5%
	High	39	21	3341	1.8%
	Average	3528	3247	3426	4.8%
	School buildings				
	Low	286	9	3	4.0%
	Medium	18	279	2	6.7%
	High	7	0	292	2.3%
	Average	311	288	297	4.4%
	Other non-residential buildings				
	Low	434	7	5	2.7%
	Medium	29	414	4	7.4%
	High	12	0	435	2.7%
	Average	475	421	444	4.3%

Learning curves were generated for the XGBoost models to analyze the balance between bias and variance in indoor radon modeling across different building classes, as depicted in Figure 3.11. The results indicated a marked reduction in logarithmic loss (logloss) when the number of decision trees ranged from 20 to 25 in the models, applicable across training, cross-validation, and validation phases. The diminution of logloss persisted until reaching a minimum, with the optimal number of trees indicated by green lines, approximately between 110 and 155 trees. The validation and cross-validation curves displayed a high degree of congruence. However, a discernible divergence between the training and validation curves suggested a potential risk of overfitting as the number of pruned trees increased. This pattern of learning curve behavior was consistently observed across all building class subgroups.

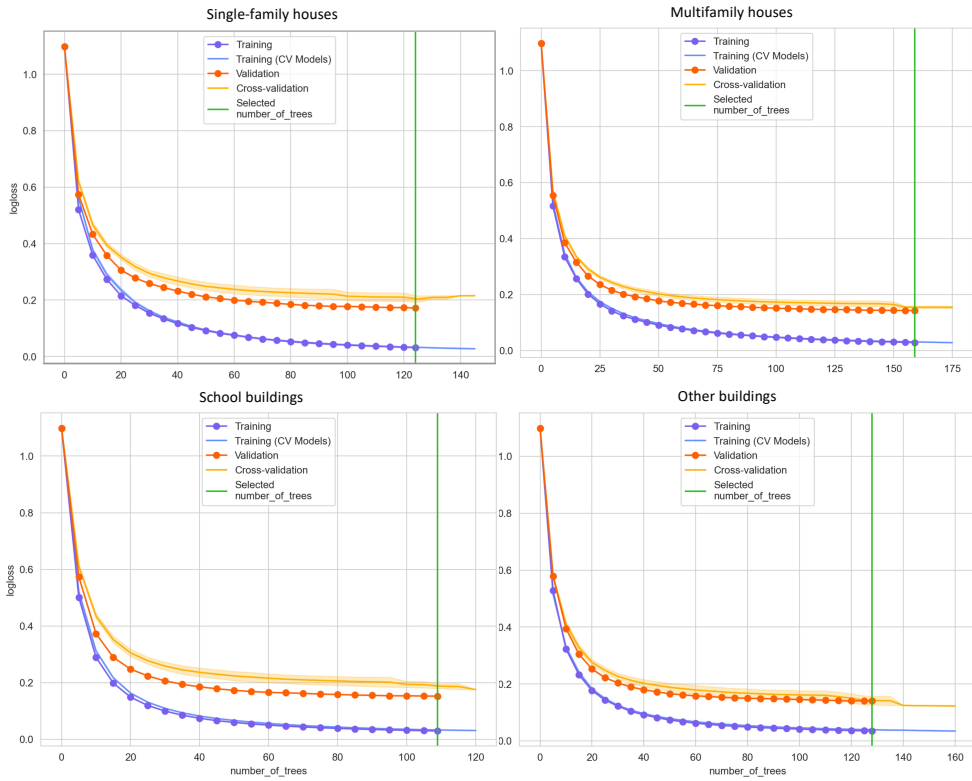


Figure 3.11. The learning curve for indoor radon interval prediction by building classes (data sourced from Paper VI).

In Figure 3.12, the significance of various features in predictive models tailored for distinct building categories is delineated, thereby facilitating model interpretation. This analysis reveals a consistent trend in the variables that exert influence in both XGBoost and DNN models within identical building classifications. For single-family houses, features such as construction year, natural ventilation, and building physical footprint were found to be crucial in estimating intervals of indoor radon levels. Conversely, multifamily houses displayed a similar hierarchical importance of features, with construction year, building physical footprint, total floor area, and the presence of exhaust ventilation systems being deemed important features. In the case of school buildings, the primary determinants identified included the physical building footprint, floor area, and the presence of basements. A more varied set of influencing factors was observed in other non-residential buildings, including aspects such as floor area, building physical footprint, geographical adjustment factors, the presence of basements, geographical latitude, construction year, and the inclusion of exhaust ventilation systems.

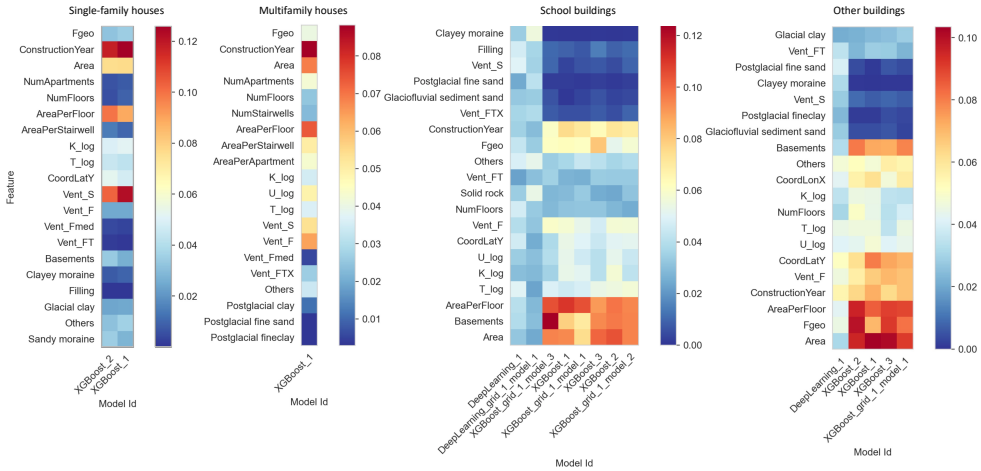


Figure 3.12. Feature importance for indoor radon interval prediction by building classes (data sourced from Paper VI).

In Appendix III, partial dependence plots (PDPs) is shown to delineate the marginal effect of key variables on indoor radon levels. These plots illustrate the impact of each variable by measuring changes in mean responses. Individual PDP plots were specifically tailored for different building classes, delineating the relationship between the target function and selected features. In these plots, the influence of each feature is represented distinctly by green lines for XGBoost models and black lines for DNN models. These findings offer a nuanced understanding of how various factors influence indoor radon levels across different building classes.

- **Single-family houses**: An increased level of indoor radon was particularly observed in single-family houses constructed around the 1960s.
- **Multifamily houses**: Buildings built between 1945 and 1980 exhibited increased indoor radon levels. Additionally, multifamily houses with a building physical footprint exceeding 2000 m² and located at latitudes above 58° showed a higher likelihood of high indoor radon levels.
- **School buildings**: The impact of construction year on indoor radon levels in school buildings was relatively subtle, displaying no distinct patterns. However, school buildings with physical footprints around 800 m² or floor areas between 2000 m² to 5000 m² were prone to higher indoor radon levels.
- **Other non-residential buildings**: For other non-residential building types (excluding school buildings), those buildings situated in the areas with geographical adjustment factors ranging from 0.9 to 1.4 were more likely to experience medium levels of indoor radon.

3.4.2. Radioactive Substances Prediction in the Building Stock

To enhance the generalizability of the model given the limited size of the training dataset, building registries from the five sampled municipalities – Stockholm, Gothenburg, Malmö, Gävle, and Umeå – were utilized. As detailed in Table 3.5, the training dataset comprised approximately 2% of the regional building stock constructed between 1930 and 1980, predominantly featuring single-family houses. This distribution suggests a potential bias in the Bayesian network models, where overrepresented building classes could disproportionately influence the learning of Bayesian networks from the data. Consequently, this bias might skew the models developed through structural and parameter learning, resulting in improved generalization for single-family houses. However, the allocation of the training dataset across different building classes was relatively balanced, with the exception of other non-residential buildings, which appeared to be underrepresented. Addressing this imbalance is crucial for achieving a representative training set that accurately reflects the diversity of the regional building stock.

Table 3.5. Statistics of regional building stock (1930-1980) for radioactive concrete estimation (data sourced from Paper V).

Building class	Single-family house	Multifamily house	School building	Other non-residential building	Total
Training size (N)	1,841	312	101	170	2,424
Buildings (N)	83,998	17,634	13,026	1,338	115,996
Training share (%)	2	2	1	13	2

In the Bayesian network models, the estimation of joint probabilistic distributions (where the sum of probabilities for all statuses of a variable equals 100%) of radioactive concrete for individual variables was presented in Table 3.6. Combining multiple variables to query specific statuses is also possible to yield probability estimation. The findings showed that approximately 33.7% of the regional building stock is likely to contain radioactive concrete in building components. In particular, buildings located within a 600 km radius from radioactive concrete manufacturing plants exhibited over tenfold increase in the likelihood of containing radioactive concrete compared to those outside this zone. Furthermore, the data indicated that around 59% of buildings constructed between 1960 and 1968 had a higher propensity for containing radioactive concrete, compared to those built between 1975 and 1980, which presented a markedly lower risk at only 2%. The likelihood of radioactive concrete presence was observed to increase with large floor area. Approximately 35% of buildings spanning between 360-1500 m² and 50% of those with floor area above 1500 m² were estimated to potentially contain radioactive concrete. Among various building types, multifamily houses exhibited the highest probability, at 19.6%, which is more than double that

of other non-residential buildings and fivefold higher than that of single-family houses and school buildings. While the existence of basements suggested a slightly increased probability, it was not identified as a direct indicator of radioactive concrete presence.

Table 3.6. Predicted joint conditional probability distribution of radioactive concrete in regional building stock of studied municipalities, as detailed in Paper V.

Variable	Status	Radioactive concrete	
		Positive	Negative
Average	-	33.7%	66.3%
Average distance (km)	below 300	13.8%	8.0%
	300-600	15.1%	3.1%
	above 600	1.2%	58.9%
Construction year	1930-1955	4.8%	16.3%
	1955-1960	7.6%	11.6%
	1960-1968	15.6%	10.8%
	1968-1975	5.7%	12.4%
	1975-1980	2.0%	13.3%
Floor area (m^2)	below 150	2.6%	13.2%
	150-220	2.7%	13.7%
	220-360	5.0%	13.6%
	360-1500	9.6%	15.1%
	above 1500	12.8%	12.0%
Building class	Single-family house	4.1%	28.6%
	Multifamily house	19.7%	16.7%
	School building	3.4%	9.7%
	Other non-residential building	8.6%	9.3%
Basements	No	11.7%	22.9%
	Yes	24.3%	41.3%

An similar approach for ensuring data representativeness was applied to the indoor radon training dataset, which featured a larger dataset size and an expanded building stock in the prediction set. Table 3.7 details the statistics for building stocks constructed between 1930 and 1980 in each Swedish metropolitan area based on valid EPC in 2023, compared with the national building stock. The analysis revealed a predominance of data from the Stockholm region in the training set, followed by considerable contributions from the Gothenburg and Malmö regions. Noteworthily, a higher frequency of indoor radon measurements was observed in multifamily houses and school buildings, approximately 2.5 times that of single-family houses and other non-residential buildings. This pattern of indoor radon measurement distribution was consistent across the metropolitan regions. On average, the training set encompassed 25% of properties in Swedish metropolitan areas, corresponding to 8.2% of the entire national building stock.

Table 3.7. Statistics of Swedish metropolitan building stock (1930-1980) for indoor radon prediction (data sourced from Paper VI).

Metropolitan*	Building class	Training size (N)	Training share (%)	Buildings (N)	Building share (%)
Stockholm region	Single-family house	20,769	27	77,071	16
	Multifamily house	18,819	59	31,866	21
	School building	1,820	54	3,382	7
	Other building	1,977	24	8,405	43
Gothenburg region	Single-family house	4,014	9	44,808	9
	Multifamily house	4,685	35	13,301	9
	School building	695	39	1,773	3
	Other building	670	15	4,382	22
Malmö region	Single-family house	2,580	8	33,082	7
	Multifamily house	2,076	20	10,175	7
	School building	324	26	1,247	2
	Other building	310	9	3,420	17
Total	Single-family house	27,363	18	154,961	33
	Multifamily house	25,580	46	55,342	37
	School building	2,839	44	6,402	13
	Other building	2,957	18	16,207	82
	Sum	58,739	25	232,912	33
Sweden	Single-family house	-	-	474,600	100
	Multifamily house	-	-	150,405	100
	School building	-	-	50,660	100
	Other building	-	-	19,731	100

*Swedish metropolitans imply the most populated areas in the Stockholm region, Gothenburg region, and Malmö region with a total of 51 municipalities.

The lead XGBoost models were utilized on property registers across Swedish metropolitan regions to estimate the proportion and characteristics of the building stock susceptible to high indoor radon levels. To assess model reliability, two datasets were prepared: properties with recorded indoor radon measurements and properties lacking such measurements, categorized by building classes for prediction purposes. The prediction results for properties with measurements were contrastive with the actual distribution of indoor radon levels as per statistical records. It was observed that the models tended to misclassify 1-3% of medium and high indoor radon intervals as low. Despite this underestimation, the predicted distributions closely mirrored the actual ones. Statistical data indicated that 9-13% of the building stock in the Stockholm region has indoor radon level higher than the reference limit, whereas in the Gothenburg and Malmö regions, only 2-6% have. A exception was multifamily houses in Gothenburg, where 12% exhibited medium to high indoor radon levels.

After evaluating model uncertainty, these models were applied to the set of properties without indoor radon measurements. As anticipated, the prevalence of

medium and high indoor radon levels in these predictions was lower than in the statistical data, likely due to the assumption that buildings with existing indoor radon measurements were suspected to have higher indoor radon exposure. The predicted shares and statistical data were then normalized based on data sizes across building classes and metropolitan regions, as depicted in Figure 3.13. This graph illustrates diverse rates of indoor radon measurement, with 40-80% of building stock in the Stockholm region having been measured, in contrast to only 18-40% in the Gothenburg and Malmö regions, particularly among single-family houses, multifamily houses, and other non-residential buildings. The substantial number of unmeasured buildings could contribute to lower model accuracy in predicting specific building categories, likely due to their limited representation in the training dataset, thereby explaining the underestimation of medium and high indoor radon levels in the predicted shares for Gothenburg and Malmö.

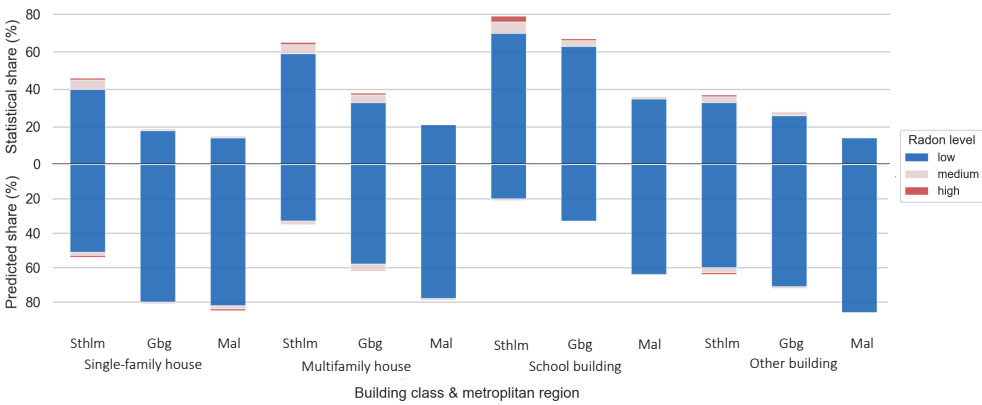


Figure 3.13. Normalized indoor radon interval distribution in the Swedish metropolitan building stock, as detailed in Paper VI.

4. Discussion

This section discusses the complexities associated with data transformation and matching, representativeness assessment of empirical data, and methodological benefits and limitations. Afterwards, the findings on the detection records of hazardous substances, prediction performance of models, and identified patterns were compared with the results in former studies, highlighting both alignments and deviations from previous studies. Finally, the chapter concludes with an overarching discussion that extends beyond the technical findings, reflecting on the broader scientific and societal contributions of the study.

4.1 Data and Methodological Limitations

The study primarily relies on empirical data from environmental inventories and indoor radon measurements to develop a data-driven approach for assessing contamination in existing buildings. The effectiveness and scope of the predictive model are substantially influenced by where and how the data were collected, while the model performance and generalizability were closely associated with the quality and timeliness of sample observations in relation to the population. The challenges addressed in the thesis could be summarized into three interrelated aspects: (1) assembling data from diverse inspection and measurement records spanning wide geographical and temporal ranges, (2) accounting for the dynamic nature of building stocks, which includes ongoing developments, demolitions, and regional differences in building typologies, and (3) the complexities associated with real-world validation at an urban scale. These challenges highlight the inherent heterogeneity and evolving characteristics of building stocks, underscoring the need for comprehensive and high-quality data as a foundation for predictive modeling.

In response to these challenges, the thesis focuses on buildings built between 1930 and 1980, a period prior to the regulation of hazardous materials. This selection targets buildings situated in the most populated areas to be representative of the general Swedish building stock in terms of the amount of buildings. The analysis and modeling are then refined based on building categories or classes, facilitating approximate inference with more closely matched instances. Given the absence of a central, digital repository for building environmental inventories, considerable effort was dedicated to accurate data transformation and the cleaning

of invalid or incomplete inventory records. The thesis details the issues encountered during data processing and method development, offering insights and recommendations for future improvements in this field.

4.1.1. Complexity in Data Transformation and Matching

The iterative process of transforming data from building environmental inventories into digital datasets forms a significant part of this thesis. The lack of a standard format for inventories and mandated quality control lead to considerable variation in the extent and quality of inventories across municipalities. This variation introduces uncertainties in interpreting the documentation, especially when typical statements for instance, “did not inspect”, “did not detect”, or “did not exist,” fail to clarify the presence of specific hazardous materials. This ambiguity primarily compromises the quality and quantity of available detection records and was partially improved by introducing supplementary information requirements in existing pre-demolition audits; such as the knowledge and experience level or certification of auditors, documents on previously executed pre-demolition audits, the choice and representativeness of samples, and post-audit of hazardous waste after demolition or renovation activities, etc. (Pereira et al., 2021). Mangold et al. (2015) and Pasichnyi et al. (2019) encountered similar data quality challenges. The former improved estimation of heated floor areas (A_{temp}) using regression analysis, while the latter proposed validation levels for EPC data quality by adding or modifying EPC variables. In this way, uncertainty associated with the inspection and diagnostic records of in-situ hazardous materials could be better assessed and quantified. This problem of data incompletenes was less prevalent in the analysis of indoor radon measurements, owing to the availability of a comprehensive digital data in the EPC register.

Retrieving data from inventories posed another challenge due to the often limited descriptions of inspected buildings or their components in inventory documents. Key details such as building address, construction year, renovation year, floor area, and number of floors are crucial for accurate spatial matching between building-specific data from empirical sources, i.e., environmental inventories and indoor radon measurements, and generic national building registers from the building database (Johansson et al., 2017). Such complementary information is sometimes insufficient or missing, making data coupling of historical inspection or measurement records with up-to-date building registers uncertain. Issues in data merging include unmatched building parameters possibly due to renovation, inadequate information from inspected buildings for proper register allocation, elimination of building registers post-demolition, discrepancies between the number of inventoried and registered buildings, and aggregated inventories from multiple buildings. Currently, multiple building entries from registers are retrieved using matched real estate indices, with missing cadastral and building variable data filled

in through backward data searches on Google Street View and hitta.se's plot map (a Swedish search engine that offers telephone directory, addresses and maps). This manual process is time-consuming and impractical for large-scale digitization of inventories across numerous municipalities.

To overcome the aforementioned limitations in data quality, quantity, and merging, establishing a uniform digital repository for building environmental data, akin to the existing EPC register, is necessary. This repository would be used to consolidate various digital inspection protocols, including pre-demolition audit inventories, indoor radon measurements, and building defect diagnostics. Pereira et al. (2020) proposed a similar global standard for building inspections and diagnostics to ensure uniform level of details in classification. Integrating the proposed repository with a digital pipeline for automated data collection, matching, and processing would enhance modeling capabilities. Providing inspectors with access to building database queries and permit documents during desk studies and inventories would enable more thorough, hypothesis-driven pre-demolition audits or indoor radon measurements. Furthermore, considering the dynamic nature of building stocks, linking registers from building databases of authorities and municipalities is essential to integrate observation units and synchronize data updates. Despite legal challenges, such integration could prevent the formation of database silos and outdated registers, facilitating accurate data retrieval and coupling.

4.1.2. Data Representativeness of Empirical Data

Representativeness of training sets is critical for assuring model generalizability across the building stock (Clemmensen & Kjærsgaard, 2022; Schat et al., 2020). The training set were analyzed to define the application scope, sample sizes and their representative populations. The hazardous material dataset contained smaller fractions of observations from several municipalities and had prediction sets at the local scale, while the indoor radon dataset had rather large nationwide samples but primarily from large municipalities and had a prediction set at the regional scale. Statistical distributions of critical building parameters, such as construction year, floor area, and building physical footprint, were consistently compared between the training and prediction sets among datasets (Clemmensen & Kjærsgaard, 2022).

The findings indicate that while the building physical footprints were similar across datasets, buildings constructed during 1945-1964 (Folkhemmet, The People's Home era) and 1965-1974 (the Million Homes Programme) were disproportionately represented. This overrepresentation stems from the data collection method's inherent bias, focusing on renovated or demolished buildings, which are more frequently audited. Consequently, large, complex, and public buildings, for which the environmental inventory quality is higher, represented a small portion of the building stock. This led to a skewed dataset, particularly in some

building categories. Residential buildings, comprising 36% of the training set, contrasted with their actual 91% representation in terms of number of buildings in the Swedish building stock (Statistics Sweden, 2023). A similar issue emerged in the analyses of the radioactive concrete dataset due to limited access to data, resulting in constrained variable intervals, such as distances to radioactive concrete manufacturing factories. Analyses of the indoor radon dataset, benefiting from extensive samples from EPC data and municipality open APIs, encountered fewer of the aforementioned issues.

There are multiple methods to estimate and correct selection bias, such as two-step statistical resampling according to population distribution followed by adjusting label proportion, or bootstrapping bias estimates and correcting sample estimates with replacement. Nonetheless, modifying the training set's sample distribution could introduce biased labels into smaller subgroups during data replication, while eliminating oversampled subgroups might reduce the training set size. Additionally, computing percentile bootstrap estimates from a single building feature, such as construction year, may not be effective in binary classification tasks. To address this issue, the dataset was partitioned based on building category or class, ensuring minimum data size in each subgroup and creating predictive models accordingly. This approach aimed to obtain representative samples as a microcosm of the target population, relevant for drawing historical inferences or making predictions for the majority (Clemmensen & Kjærsgaard, 2022). To enhance building stock coverage and robustness against distribution shifts, expanding the data pool in size and geographic representation is essential for future environmental inventory collection.

4.1.3. Methodological Benefits and Limitations

Compared to other data-driven methods for detecting in situ hazardous substances, this proposed approach offers several advantages. The first is its methodological comprehensiveness, encompassing a wide range of prediction targets and building classes. Unlike prior studies focusing solely on asbestos or PCB, this approach is not limited to specific hazardous materials, as it utilizes environmental inventories documenting multiple materials. The inventories, randomly collected from various building classes renovated or demolished in the past decade, reflect the diverse building stock and the actual presence of hazardous materials.

Another significant benefit is the method's flexibility, allowing for easy adaptation to different regional or national building stocks. Given that building stocks are inherently dynamic and analytical results can quickly become outdated, it is crucial to develop models capable of frequent updates. As additional data from other municipalities become available, new inventory or measurement data can easily be integrated into existing models, enabling continuous training. This

adaptability also allows models developed for one hazardous material to be swiftly applied to others.

However, certain methodological limitations must be acknowledged. Misclassification errors in some labels persist, and model performance could be enhanced by increasing training set size, minority class representation, and feature availability. The precision of binary predictions depends on the data volume of labels from each class. Incorporating underrepresented observations improves data representativeness, model granularity, and accuracy. Missing critical features in registers, such as building materials, renovation years, and historical hazardous material manufacturing locations for PCB and asbestos, or foundation types and radioactive concrete presence for indoor radon prediction, can limit classifier effectiveness. Variable renewal rates in registers also pose a risk of data mismatches for older inventories. Consequently, environmental inventories from the last decade (2010-2022) were prioritized to align with EPC register updates.

To implement these data-driven approaches, some prerequisites are necessary. Identifying a common key for data coupling between environmental data and building registers is crucial. Without it, predictive analysis is hindered by incomplete data and high missing value counts. Previous studies, limited to descriptive analysis and statistical inference, often lacked access to building databases and identifiers. Training models on small datasets risks overfitting, while training on heterogeneous samples can lead to underfitting and poor performance. Therefore, data partitioning must balance similarity in building characteristics with adequate subgroup data size. With increasing public awareness of contaminant exposure and material circularity, the prevalence of environmental inventories and indoor radon measurements is growing. Countries including Canada, Poland, France, and others have established national databases for asbestos and/or PCB inventory. This expanding data pool lays a solid foundation for replicating the data-driven approach in other contexts and developing new digital data collection and processing pipelines.

4.2 Result Implications

The section outlines the overall implications of the study's results, ranging from statistical descriptions to model performance and pattern interpretation. This investigation presents Sweden's first comprehensive statistical and predictive analysis into asbestos, PCB, and radioactive concrete-containing materials. Comparing the thesis findings with previous nationwide surveys and literature is essential for identifying potential dataset skewness and verifying the reliability of the training data. Prior research on in situ hazardous materials largely focused on quantity estimation and mapping, whereas this study concentrates on predicting the probability distribution and characterization of these materials. Limited previous

studies, whose statistics are useful for benchmarking, enable comparison of hazardous material detection rates among building stocks and evaluation of the predictive performance of various modeling approaches. The study leverages feature correlations identified in previous research (outlined in section 3.1.1), expert assumptions from industrial inputs (detailed in section 3.1.2), and descriptive statistics from this research (presented in Papers IV-VI) to validate the prediction results and discern patterns of hazardous substances. This comprehensive approach enhances the relevance of the study's findings.

4.2.1. Statistics of Hazardous Substance Records

The PCB statistics derived from environmental inventories in Swedish municipalities revealed the average detection rates (47%) slightly higher than those reported in the former BETSI survey (Swedish National Board of Housing Building and Planning (2010), which estimated that at least 34% of Swedish buildings constructed or renovated between 1956 and 1973 contained PCB. This discrepancy might be due to incomplete PCB inventory during 2007-2008 in the BETSI survey, suggesting that the actual number of buildings with PCB could be higher. This study indicated that capacitors were the most frequently detected PCB-containing materials (51%), followed by joints or sealants (21%), as detailed in Table 4.1. Acrylic flooring had the lowest detection rate (3%). However, the detection of PCB in paint and plaster remained unclear due to few records in the environmental inventories. To determine PCB detection rates in door closers and oil-containing cables, their data sizes also need expansion.

Table 4.1. Detection rates of PCB-containing materials in buildings.

Building	**Robsen et al. (2010) / Diamond et al. (2010)**	**Paper IV**
Location	Toronto, Canada	Sweden
Year built	1945-1980	1930-1985
Number	455	786
Category	All buildings except single-family houses	All buildings
Material	**Available record (N) (%positive detection)**	
Joints or sealants	455 (14%)	406(21%)
Double-glazed sealed window	-	336(17%)
Capacitors	-	291(51%)
Acrylic flooring	-	271(3%)
Door closer	-	78(42%)
Cable with oil	-	95(16%)

These findings are consistent with existing literature, which identifies transformers and building sealants as the primary sources of PCB legacy (Shanahan et al., 2015), particularly in schools, hospitals, and downtown commercial buildings (Diamond et al., 2010). Previous studies, however, provided limited data on local PCB detection rates per building material, indicating a need for more comprehensive measurements of total mass in buildings and concentrations in specific components. For instance, it was assumed that around 14% of PCB-containing joints and sealants were present in buildings built between 1945 and 1980 in Toronto (Diamond et al., 2010; Robson et al., 2010a), but the prevalence of PCB in other building components was not thoroughly investigated. Bergsdal et al. (2014) estimated that Norwegian non-residential and residential building stocks contained 231 tonnes and 156 tonnes of PCB, respectively. This estimate included 82 tonnes in joints, 95.5 in double-glazed sealed windows, 124.9 tonnes in capacitors and lighting fixtures, 82 tonnes in plaster, and 2.8 tonnes in paint. These estimates provide insight into the accumulated mass and suggest the prevalence of various PCB materials, but they do not directly translate into detection rates and thus are not included in Table 4.1. The need for more detailed and comprehensive data to accurately assess PCB presence in building materials is evident.

Contrastingly, in situ asbestos materials in dwellings were more comprehensively characterized in terms of their presence and quantity, as detailed in Table 4.2. The study revealed that asbestos was detected in 78% of the Swedish building stock. This finding aligns with the BETSI survey by the (Swedish National Board of Housing Building and Planning (2010), which reported asbestos in 40% of single-family houses and 50% of multifamily houses, particularly those built between 1961 and 1975 (50% in single-family houses, 75% in multifamily houses) and before 1960 (56% in single-family houses, 69% in multifamily houses).

The most prevalent sources of asbestos in the Swedish building stock were pipe insulation (65%), door or window insulation (61%), and cement panels (60%). The secondary sources included floor mats (49%), joints or sealants (49%), ventilation channels (42%), and floor mat glue (40%). These detection rates, especially in parts of the building stock from the Million Homes Programme, were higher than those reported in other countries. For instance, Franzblau et al. (2020) found a 95% detection rate in abandoned residential buildings in the City of Michigan, while Govorko et al. (2019) concluded that 82% of Australian houses contained asbestos, frequently found in floor mats, door or window insulation, and cement panels (Franzblau et al., 2020; Govorko et al., 2019; Krówczyńska et al., 2020). Song et al. (2016) reported that 85% of buildings they studied contained asbestos, with 73% of samples containing asbestos materials, predominantly in ceiling installations. However, certain asbestos materials, such as tile or clinker, floor mat glue, joint or sealant, and valve, were not explored in previous studies, making direct comparisons impossible. The comprehensive nature of this study's findings, covering a broad spectrum of asbestos-containing materials, underscores its significance in understanding asbestos prevalence in building stocks.

Table 4.2. Detection rates of asbestos-containing materials in buildings.

Building	Franzblau et al. (2020)	Govorko et al. (2019)	Krówczyńska et al. (2020)	Paper IV
Location	Michigan, US	Australia	Poland	Sweden
Year built	1885-	1985-1990s	N/A	1930-1985
Number	605	702	6,287	786
Category	Residential dwellings	Residential dwellings	All buildings	All buildings
Material	**Available record (N) (%positive detection)**			
Pipe insulation	115 (19%)	77 (11%)	-	468 (65%)
Door/wind. insulation	181 (30%)	-	-	372 (61%)
	171 (28%)	-	-	-
	135 (22%)	-	-	-
Cement panel	291 (48%)	310 (44%)	2892 (46%)	322 (60%)
	204 (34%)	126 (18%)	-	-
	136 (22%)	68 (10%)	-	-
Tile or clinker	-	-	-	455 (36%)
Floor mat glue	-	-	-	376 (40%)
Floor mat	310 (51%)	187 (27%)	-	365 (49%)
Vent. channel	192 (2%)	-	-	310 (42%)
Joint/sealant	-	-	-	261 (49%)
Switchboard	-	351 (50%)	-	71 (27%)
Valve	-	-	-	144 (35%)

Radioactive concrete, while rare in other countries, was extensively used in Swedish construction from 1929 to 1975. Early investigations of its prevalence in the Swedish building stock included gamma radiation scanning by vehicles in the 1980s, the ELIB survey in the 1990s (Sedin & Hjelte, 2004), and the BETSI survey in 2010 (Swedish National Board of Housing Building and Planning, 2010). The ELIB survey found that radioactive concrete resulted in an increase of the average indoor radon level by 10% in single-family houses and 20% in multifamily houses, based on comparisons between 42 buildings with radioactive materials and 672 without. Khan et al. (2021) reported a 63% increase in average indoor radon levels in residential dwellings built with radioactive concrete, which is lower than the findings of this study that indicate two to four times higher indoor radon concentration in 398 residential buildings with radioactive concrete compared to 1,808 buildings without. Additionally, the study revealed that about 18% of the Swedish building stock constructed between 1930-1980 contained radioactive concrete, surpassing the previous estimate of 6-7% for buildings built before 2005, as identified in the BETSI survey. The likelihood of encountering radioactive concrete was highest in buildings erected between 1960 and 1968, corroborating findings from the BETSI survey that the maximum likelihood of radioactive concrete presence was in buildings built between 1961 and 1975, and before 1960.

On the other hand, nationwide indoor radon surveys in Sweden, conducted nearly every decade, provided accessible statistics on annual average concentrations and the proportion of buildings with indoor radon levels above the reference level. The shares of buildings exceeding the reference level were similar across different studies: 11% in the ELIB survey, 10% in the latest report from the Swedish Radiation Safety Authority (SSM), and 12% as reported in this thesis. The annual average indoor radon concentration of 110 Bq/m^3 found in this study aligns with the ELIB survey's 108 Bq/m^3. However, the average indoor radon levels in single-family houses were slightly lower than those in previous surveys, while the levels in multifamily houses were higher. This discrepancy could be attributed to the selection of building stock from 1930-1980 for this study, a period characterized by extensive use of radioactive concrete in major housing production programs, leading to generally higher indoor radon concentrations.

Table 4.3. Statistics of radioactive substances from Swedish indoor radon surveys.

Radon survey	**ELIB survey**	**Radon Survey**	**BETSI survey**	**SSM report**	**Papers V-VI**
Time	1991/1992	2000s	2007/2008	2007/2009	2023
Data size	1,360	215,000	1800	387,347	114,857
Construction year	-1988	-	-2005	-2020	1930-80
Building category	Residential buildings	Residential, schools, care homes	Residential buildings	All buildings	All buildings
	Count of building (N) (%share above 200 Bq/m^3)				
Single-family house	714 (17%)	215,000 (35%)	(14%)	340,000 (16%)	53,533 (14%)
Multifamily house	646 (7%)	44,200 (28%)	(7%)	440,000 (17%)	49,139 (12%)
School building	-	-	-	1,005 (-%)	5,660 (9%)
Other building	-	-	-	2,342 (19%)	6,525 (9%)
Total	1,360 (11%)	-	-	387,347 (10%)	114,857 (12%)
	Annual average indoor radon concentration (Bq/m^3)				
Single-family house	141	-	124	128-136	118
Multifamily house	75	-	-	79	105
School building	-	-	-	105	98
Other building	-	-	-	106	105
Total	108	-	-	-	110

4.2.2. Performance Evaluation of Prediction Models

The study explored statistical approaches, machine learning models, and neural networks to evaluate their methodological strengths and weaknesses in predicting hazardous substances. It was found that Bayesian network and tree ensemble classifiers, such as DRF/XRT, GBM, XGBoost, and CatBoost, were more effective than neural networks for estimating the probability distribution of hazardous materials using tabular environmental inventory data. Bayesian networks offered superior model explainability, particularly in understanding dependencies and causal relationships between variables and in providing conditional probability distributions under various evidence combinations. Machine learning classifiers, on the other hand, excelled in prediction granularity and accuracy for individual buildings, handling missing values and high-dimensional datasets. Both types of models were adaptable to new data inputs and efficient in anomaly detection.

The model performance in this study was compared with other research in Table 4.4. Due to a lack of literature on PCB materials and radioactive concrete prediction, the model evaluation focused on asbestos-containing building materials. The machine learning models developed in this study achieved higher performance (AUC = 0.61-0.93 for residential buildings, AUC = 0.65-0.85 for non-residential buildings) compared to the random forest regression models (Pseudo-R^2 = 0.76 for all buildings) by Wilk et al. (2019), which used similar feature sets but smaller datasets for predicting asbestos cement panels. To further minimize logarithmic loss, incorporating additional variables such as distance to historical hazardous material manufacturing plants and material-specific features could be beneficial. Understanding error types and introducing penalties in cost-sensitive learning, as well as increasing the training dataset size, are essential for future model refinement.

Krówczyńska et al. (2020) and Raczko et al. (2022) used CNN models for recognizing asbestos-cement roofing in Poland, achieving 88-93% overall accuracy. However, their approach, limited to image data and a single material type, lacked the flexibility of the method proposed in this study, which is capable of categorizing various indoor, built-in, or non-visual hazardous materials and utilizing existing pre-demolition audit data. This distinction highlights the adaptability and broader applicability of the proposed method in hazardous substance prediction.

Table 4.4. Comparison of model performance for asbestos material prediction.

Dataset specification	**Wilk et al. (2019)**	**Krówczyńska et al. (2020)/ Raczko et al. (2022)**	**Paper IV**	
Application	Map and quantify asbestos-cement products	Identify asbestos-cement roofing on aerial images	Predict probability of asbestos materials and their spatial distribution	
Input data	Field inventory of asbestos-cement roof	Aerial image of asbestos roofing	Environmental inventory	
Data size	6,287	3,124	287	499
Algorithm	RF regression	CNN image recognition	LR, SVM, *k*-NN, GLM, DRF/XRT, GBM, XGBoost, DNN, Stacked ensemble classification	
Feature	Social-economic data, building features (roof slope, type), localization (manufacturing plants)	-	Postcode, building category, building type, construction year, floor area, numbers of floors, basements, ventilation types, building physical footprint	
Label	Asbestos-cement products (type, quality, amount)	Image signature of building roof (type of roofing, degree of roof pitch)	Asbestos components (detection, quantity)	
Category	All buildings	All buildings	Residential	Non-Residential

Material	**Performance metrics (1e-2)**			
	Pseudo-R^2	**Overall accuracy**	**AUC**	
Pipe insulation	-	-	81	75
Door/win.insul.	-	-	93	85
Cement panel	76	88-93	84	65
Tile or clinker	-	-	83	69
Floor mat glue	-	-	88	73
Floor mat	-	-	61	68
Vent. channel	-	-	90	74
Joint/sealant	-	-	78	76

Similarly, non-parametric regression tree models and supervised classifiers were utilized to predict indoor radon concentrations and their intervals. Both MARS models and random forest models employed a hierarchical pruning approach during training to clarify feature importance and coefficients. However, these models were only partially successful in accurately fitting the complex indoor radon measurement data. The regression models, which used a simplified feature set and a smaller dataset, demonstrated lower performance compared to models developed by Kropat et al. (2015a) using kernel regression and Kropat et al. (2015b) with BART and RF. This disparity could be attributed to the different reference levels used – 300 Bq/m³ in Switzerland as opposed to 200 Bq/m³ in Sweden.

Reformulating the prediction problem into a consecutive multi-classification of indoor radon intervals yielded significant performance improvements when using XGBoost models (macro-F1 between 0.93-0.96). Despite the large size of the dataset, neural network models only achieved moderate macro-F1 scores of 0.64-0.74. The classification errors in these models might be due to several factors: the presence of unknown building conditions with radioactive concrete in some training data, the absence of key features such as foundation type, and a scarcity of observations in the higher indoor radon interval labels. These findings highlight the importance of comprehensive feature selection and the potential limitations of model performance in predicting indoor radon levels.

Table 4.5. Comparison of model performance for indoor radon prediction.

Dataset specification	**Kropat et al. (2015a)**	**Kropat et al. (2015b)**	**Conf II**	**Paper VI**
Input data	Indoor radon measurements	Indoor radon measurements	Indoor radon measurements	Indoor radon measurements
Data size	240,000	240,000	79,944	34,983
Algorithm	Kernel regression	BART*, RF regression	MARS, RF regression	XGBoost, DNN classification
Feature	Building type, foundation type, construction year, detector type, coordinates, altitude, lithology, temperature,	Building type, foundation type, construction year, detector type, coordinates, altitude, lithology, temperature,	Building paramters, radioactive susbstance, soil type	Building paramters, radioactive susbstance, soil type
Category	All buildings	All buildings	All buildings	All buildings
Application	Predict spatial radon distribution	Predict spatial distribution of radon	Predict indoor radon level	Predict indoor radon interval
Indoor radon	**Performance metrics (1e-2)**			
	R^2	**R^2**	**R^2**	**Macro-F1**
Concentration	28	29-33	13-24	-
Intervals	-	-	-	93-96 64-74

*BART: Bayesian additive regression trees.

4.2.3. Interpretation of the Identified Patterns and Prediction

The identified patterns and predictive model results offer insights for evaluating expert assumptions and findings from previous studies. However, the presence patterns and feature impact magnitude of each hazardous material varied across building categories, making it challenging to generalize the findings due to regional building stock and building class differences. Although the model performance for residential buildings was higher than for non-residential buildings, the smaller data size and less pronounced feature impact in residential buildings should be considered when assessing prediction robustness.

Distinctive patterns of PCB capacitors were noted in the existing building stock. Early-built non-residential buildings, except schools, lacking balanced ventilation with heat exchangers, and large residential buildings with exhaust ventilation were more likely to contain PCB capacitors. PCB joints were more common in medium to large post-war non-residential buildings, especially commercial buildings with balanced ventilation. This aligns with findings by Diamond et al. (2010), who found in situ PCB mass to be proportional to a building's volume and electricity demand. Consequently, large downtown commercial and public infrastructure buildings had significant volumes of PCB capacitors. Non-inventoried residential buildings built between 1930 and 1980 in Stockholm, Gothenburg, Malmö, and Kiruna showed a lower likelihood (2%) of containing PCB capacitors compared to inventoried ones (36%), possibly due to PCB capacitors had already been removed during earlier renovations. However, 44% of non-inventoried non-residential buildings were estimated to contain PCB capacitors, similar to the rate in inventoried buildings (56%). Glüge et al. (2017), Robson et al. (2010a), and Shanahan et al. (2015) mapped total PCB per unit building space and city ward, but they are not comparable to this study's findings.

The presence of asbestos materials also varied, with material-specific feature impacts identifiable to varying extents, potentially useful for characterizing contaminated buildings. In residential buildings, building typology-related features (building physical footprint, number of floors, construction year, municipality) were significant, whereas construction year and location (using postcodes) were key for asbestos detection in non-residential buildings. The likelihood of finding asbestos panels was higher in medium-sized postwar buildings with larger physical footprints, consistent with Song et al. (2016), who noted an increase in asbestos detection likelihood with building age and area ratio in South Korea. Wilk et al. (2015) identified determinants for asbestos-cement roofing in Poland, including proximity to asbestos manufacturing plants and local economic conditions. However, the study's local correlation with asbestos use in buildings could not be fully explored due to data limitations. The estimated likelihood of encountering asbestos in non-inventoried residential buildings was lower (72% versus 57% for pipe insulation and 67% versus 14% for door and window insulation) compared to inventoried buildings. Conversely, the probability of detecting asbestos in non-inventoried non-

residential buildings was predicted to be higher or equal. To enhance result certainty, increasing the training sample size for residential buildings is needed to better represent the residential stock.

Average distance to radioactive concrete manufacturing plants, building class, construction year, and floor area were significant factors in the presence of radioactive concrete in the Swedish building stock. Radioactive concrete was found to be more prevalent in multifamily houses and other non-residential buildings than in school buildings and single-family houses. Especially, the probability of radioactive concrete presence increased with floor area but was not linked to the presence of basements. Specifically, radioactive concrete was most likely detected in multifamily houses with basements built during 1960-1968. These findings are consistent with current expert assumptions, such as those by Clavensjö & Åkerblom (2020), suggested that a significant portion of radioactive concrete-containing buildings were multifamily houses built during the Million Homes Programme, as well as some single-family houses and high-rise buildings in specific municipalities. Prior studies identified radioactive concrete in various building components, but its presence in non-residential buildings had not been investigated until this study, which characterized it in walls, floors, foundations, and other components in school buildings and other non-residential buildings.

The study also interpreted the interplay between features and indoor radon concentration for each building class, identifying buildings prone to high indoor radon levels. These included single-family houses built in the 1960s with natural ventilation, postwar multifamily houses above the 58° latitude with large floor areas and exhaust ventilation, and school buildings with basements and building size (including floor areas and building physical footprint) within specific intervals. Other non-residential buildings within certain geological zones (referring to latitudes) and size ranges (including floor area and building physical footprint) were also identified. Key contributing factors to indoor radon were found to be construction year, ground uranium concentration, and natural ventilation, while balanced ventilation with heat exchangers had a mitigating effect.

The general conclusions align with previous research by Olsthoorn et al. (2022) and Khan et al. (2021), highlighting the correlation between indoor radon and uranium concentration, particularly in older buildings with natural ventilation. The SSM report by Rönnqvist (2021) also noted higher indoor radon levels in older buildings and those with natural ventilation, but found no significant decrease in indoor radon levels in newly-constructed workplace buildings. This study's findings on indoor radon trends across building classes correspond with the thesis's findings. Yet the purported correlation between clay soil and indoor radon was not supported by this study, indicating its minimal impact. Moreover, previous attempts to estimate the number of dwellings exceeding indoor radon reference levels varied significantly. The SSM report's estimate was higher than the estimation in the BETSI survey but closer to the findings in the ELIB survey. This study advanced these estimates by providing more detailed predictions of indoor radon intervals for

each Swedish metropolitan region, revealing varying shares of buildings exceeding the 200 Bq/m³ threshold. The lower ratios in non-measured building stocks suggest that past measurements were more frequently conducted in buildings suspected of high indoor radon exposure.

4.3 Research Contributions

The research addresses the gaps in scalable and cost-efficient methods for comprehensive urban-wide inventories and estimation of in situ hazardous materials. Figure 4.1 in the thesis provides a synthesis of the research contributions in relation to the gaps identified in Section 1.3.2. Conf I and Paper II played a pivotal role in reviewing the usability of pre-demolition audit inventories of building environmental information and their digital transformation. By integrating building registers with environmental inventories from the past decades across four Swedish municipalities, the research enabled detailed descriptive analyses. These analyses focused on the detection rates of various PCB and asbestos components, differentiated by building classes and municipalities. Furthermore, an evaluation of the inventory types, data sizes, and proportions of missing values for each data subgroup in the constructed data matrix was conducted. This assessment led to the identification of building components with high assessment scores, earmarked for advanced predictive modeling.

Papers III and Conf II delved into the advantages and limitations of data-driven applications for predicting hazardous substances. This exploration involved preliminary development of several predictive approaches, encompassing statistical methods, machine learning, and neural network models. These models were trained and evaluated to determine the most suitable approaches for predicting hazardous materials and radioactive substances. The research also scrutinized modeling requirements, such as minimum data size, building class composition, and sample representativeness, in the context of regional building stocks.

Finally, the developed predictive models for in situ PCB, asbestos materials, radioactive concrete, and indoor radon, as detailed in Papers IV-VI, traced the presence patterns of hazardous substances by quantifying the correlation and impact magnitude of key features. The resulting probability distributions and labels of hazardous substances offer a valuable tool for contamination screening, aiding in the prioritization of comprehensive material sampling or indoor radon measurements in the building stock. This holistic approach to hazardous material prediction represents a significant advancement in addressing urban environmental challenges.

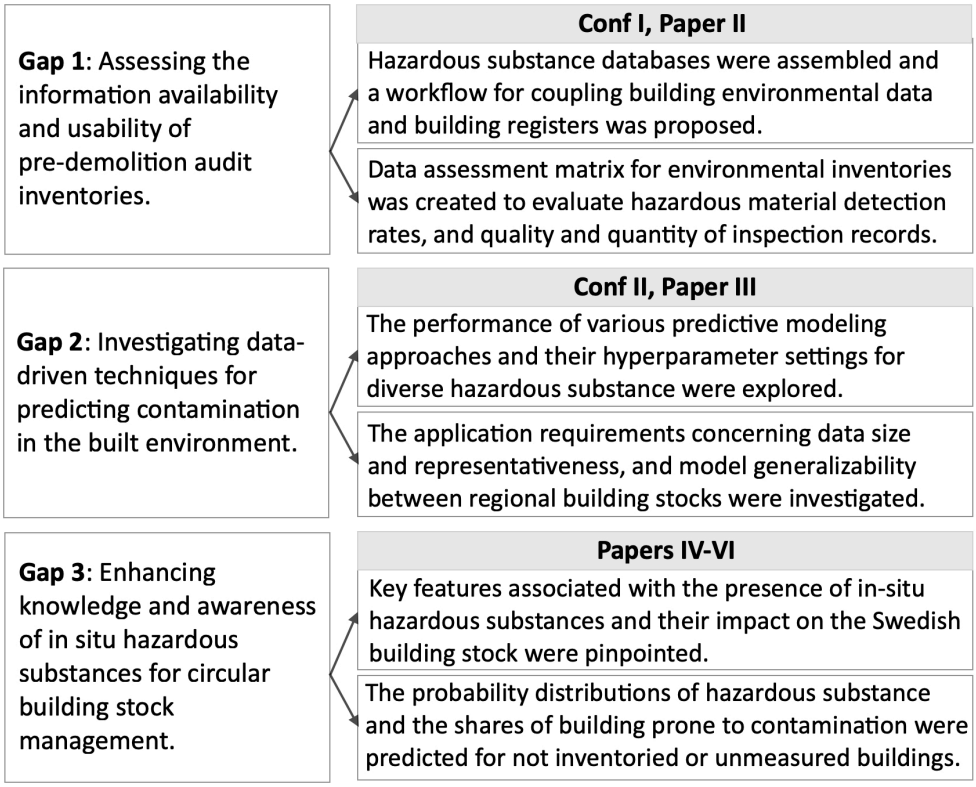


Figure 4.1. Research contributions.

4.3.1. Scientific Contributions

The thesis represents pioneering research in the utilization of information in pre-demolition audit inventories. It illustrates a workflow for analyzing and modeling existing building environmental inventories, overcoming the challenges of digitizing and merging unstructured data with building registers. The creation of hazardous material and indoor radon databases facilitated comprehensive investigations of in situ hazardous substances within the Swedish building stock. Previously, such extensive digital datasets, covering various building types and regions and linked to national building databases, did not exist, hindering the ability to trace numerous hazardous building materials on an urban scale.

In the literature, urban-wide PCB and asbestos stock inventories have been addressed using bottom-up or top-down approaches, tailored to specific substance properties. Empirical data from measurements and field sampling were utilized to estimate the asbestos and PCB in building stocks based on the estimated number and size of the buildings constructed during selected periods. Prior studies using

bottom-up approaches focused on source-centric inventory at the component and building levels, concentrating on mass and emission estimation per location (Diamond et al., 2010; Robson et al., 2010b; Tadas et al., 2011; Neitzel et al., 2020) or per sample (Franzblau et al., 2020; Govorko et al., 2019; Powell et al., 2015), and mapping their spatial distribution (Diefenbacher et al., 2016; Shanahan et al., 2015; Wilk et al., 2017, 2019). However, significant concentration variations between and within buildings (Bergsdal et al., 2014) posed challenges in generalizing results and applying them to other contexts. Conversely, top-down studies estimated the lifespan of hazardous components to assess hazardous material output and stock over extended periods, based on cumulative PCB or asbestos use data in tonnes per application and building category (Bergsdal et al., 2014; Donovan & Pickin, 2016b; Glüge et al., 2017; Zoraja et al., 2021). These approaches generally resulted in high uncertainty in lifespan estimations and aggregated findings lacking the granularity necessary to determine hazardous materials in individual buildings.

This thesis overcomes these limitations with a data-driven approach, employing descriptive analysis and predictive modeling of inspection or measurement records at the component level. The analytical results offer detailed insights into various hazardous materials and radioactive substances across different building categories or classes, age cohorts, and geographical areas. A digital toolbox was developed for predicting hazardous substances in regional and metropolitan building stocks, encompassing asbestos, PCB materials, radioactive concrete, and indoor radon. Component-specific machine learning models, optimized through cross-validation and hyperparameter tuning, demonstrated optimal performance. The patterns identified in the models' interpretations highlight critical attributes for high-resolution detection of each hazardous substance, supporting existing expert assumptions. Application of the model to regional building databases enabled both top-down analysis of building stock probability distribution and bottom-up contamination prediction for individual, previously uninvestigated buildings. These prototype models provide a cost-efficient method for screening potential hazardous substances in both local and large-scale building stocks and offer the flexibility for extension to other regions and the possibility for updates with new inspection or measurement data.

4.3.2. Societal Contributions

The research findings significantly contribute to the broader implementation of the EU Construction and Demolition Pre-demolition Audit Protocol (ECORYS, 2016) and the EU Guideline for the Waste Audits Before Demolition and Renovation Works of Buildings (European Commission, 2018), focusing on contaminated building material screening. For countries with available building-specific environmental data, i.e., environmental inventories, indoor radon measurement records, and building registers, the proposed data-driven methods can be replicated

and applied. At the national level, they support the circular transition efforts of the Swedish construction sector to achieve environmental and climate goals (Swedish National Board of Housing Building and Planning, 2023) by: (1) mapping current pre-demolition audit practices and offering managerial recommendations to enhance inventory documentation; (2) examining the usability of environmental inventory information to create a machine-readable dataset structure; (3) proposing predictive models as digital tools for identifying in situ hazardous materials and facilitating their safe, circular management. The results are presented to industry stakeholders and public authorities in a follow-up workshop, exploring practical applications of the models in current construction practices for improved hazardous material management during building retrofit and demolition.

Figure 4.2 underscores the study's contributions in relation to the existing pre-demolition audit process (left) and potential future research directions (right). A digital environmental inventory dataset, comprising historical detection records of hazardous materials integrated with building registers, was established, paving the way for predictive model development. This pilot effort lays a crucial foundation for future work in developing digital inventory protocols and data processing pipelines. Additionally, numerous machine learning models for predicting asbestos and PCB materials were developed, capable of estimating the presence of hazardous materials in buildings not yet inventoried. These models could be further incorporated into web interfaces, such as H2O Wave (realtime web applications and dashboards for AI), enabling CDW actors to estimate the probability distribution of hazardous materials in individual buildings. This would facilitate risk-based inspections and material sampling, enhancing the overall efficiency and safety of building retrofit and demolition process.

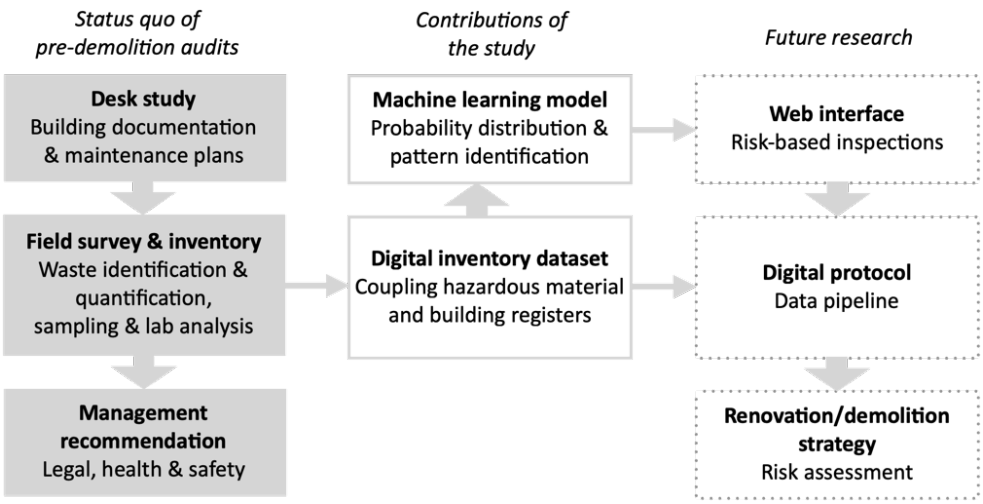


Figure 4.2. Exploitation of scientific results to the current pre-demolition audits and future prospects.

Relevant authorities responsible for housing, planning, radiation safety, and public health could benefit from contaminant screening and monitoring of buildings identified as having a high probability of contamination from a macro-perspective. Meanwhile, property owners and demolition or waste handling companies could use these results to inform their decontamination planning and hazardous waste management strategies. Enhanced knowledge of the presence of in situ hazardous substances in existing buildings is crucial for developing policy instruments for resource-efficient remediation. In addition to industry-side resource and waste guidelines for construction and demolition (Byggföretagen, 2019), the predictive outcomes could aid in risk-based inspection for semi-selective demolition, assessing the reusability of reclaimed materials and promoting quality assurance.

To fully realize the potential of this digital tool as a decision-support system for hazardous material assessment before renovation and demolition, further development of the current digital dataset and models is necessary. The primary digital infrastructure that needs establishment is a protocol for collecting and maintaining building environmental information, including data requirements and metadata criteria for auditors and authorities. Another key step is integrating the models into a web interface, enabling probability queries for non-inventoried and unmeasured buildings for public access. This would make building environmental information more readily available for desk studies in pre-demolition audits. Digital inspection records and preliminary prediction results could assist property owners in defining specifications for pre-demolition audit procurements, outlining inventory scope and field sampling requirements. Lastly, appending empirical data on the cost and time associated with material sampling and decontamination to the model delivery would enable a more accurate estimation of budgets and schedules for renovation and demolition strategy formulation.

5. Conclusions

In conclusion, this thesis is about predicting the presence of hazardous substances, including hazardous materials and radioactive substances, in the Swedish building stock at both regional and individual levels using data-driven methods. The predictive outcomes of tree-ensemble machine learning models, presented as probabilities and labels, facilitate the estimation of contamination. This probability assessment aids in prioritizing buildings for comprehensive environmental inventories or indoor radon measurements. The predictions indicate that non-inventoried and unmeasured buildings constructed between 1930 and 1980 have lower contamination levels than those assessed by existing environmental inventories and indoor radon measurements. This discrepancy is likely due to the overrepresentation of non-residential buildings and those built between the 1950s and 1970s in the inventory data. Additionally, buildings suspected of containing radioactive concrete or exhibiting excessive indoor radon levels were more frequently measured. The underrepresentation of residential buildings, which form the majority of the actual building stock in numbers, contributes to higher uncertainty in the findings.

The study also identifies a higher than average probability of hazardous material detection in medium to large-scale postwar urban buildings, which explains the lower predicted contamination exposure in the overall unaccounted building stock. To improve the accuracy and reliability of the predictive models, it is recommended to expand the training dataset. This expansion should aim to align more closely with the prediction set's composition, particularly in terms of building class, construction year, and physical building footprint. Such enhancements will refine the predictive capability, offering a more precise and comprehensive understanding of hazardous material presence in the Swedish building stock.

5.1 Concluding Findings for Each Research Question

In alignment with the research aim of utilizing data analytics and machine learning to predict and interpret the presence of in situ hazardous substances in existing building stock as a decision support for relevant actors, three research questions were formulated and addressed, with their brief answers presented in Figure 5.1. The summarized findings confirmed the effectiveness of employing supervised

statistical and machine learning learning techniques on hazardous substance presence. Environmental inventories and indoor radon measurement records served as label data, while building registers, along with geologic and geographic attributes, were utilized as training data.

This developed approach provides valuable decision support for relevant stakeholders in the building sector. It enables them to assess the extent of building contamination, thereby facilitating safer and more predictable material sorting and handling in CDW management. The application of machine learning approach on building-specific environmental data offers a significant advancement in managing hazardous substances in the built environment. This methodology not only enhances the accuracy of building component contamination assessment but also contributes to the optimization of resource management and waste reduction strategies, aligning with broader environmental and sustainability objectives.

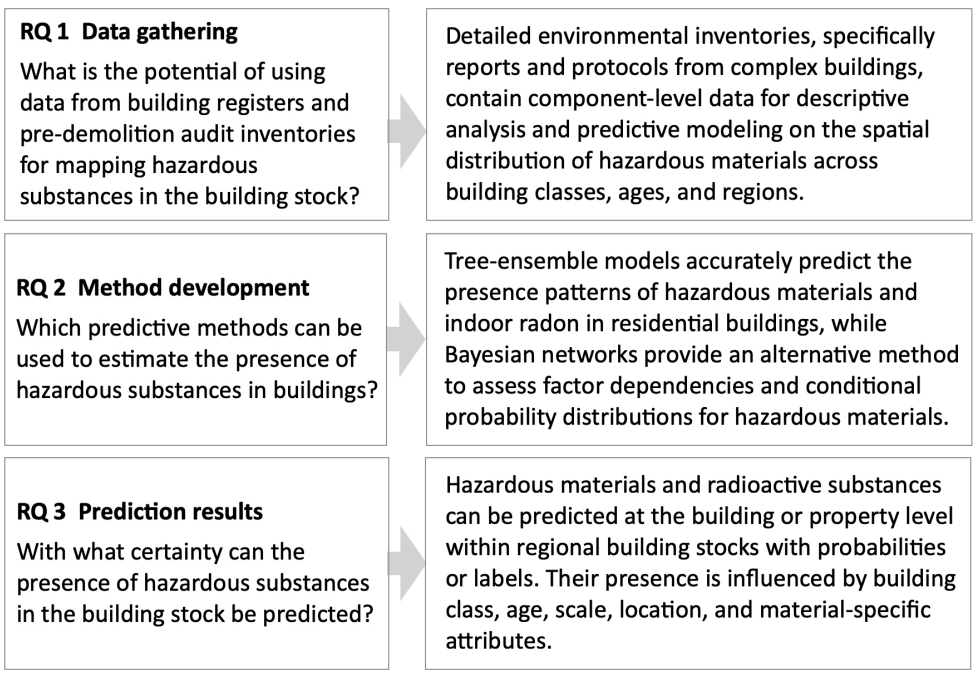


Figure 5.1. Summary of research questions and corresponding findings.

More detailed and in-depth elaboration of each answer was presented as follows:

RQ1 *What is the potential of using data from building registers and pre-demolition audit inventories for mapping hazardous substances in the building stock?*

The study involved compiling data from various types of environmental inventories with different levels of detail and inspection certainty, submitted as a basis for renovation, retrofit, re-purpose, and demolition permit applications. This data collection method provided diverse sample, yet also led to potential data skewness and selection bias, particularly the oversampling of buildings from 1945-1974, known as the People's Home and the Million Homes Programme era, and the overrepresentation of non-residential buildings compared to the Swedish building stock composition. These biases were addressed through data clustering for building category-specific modeling, as described in Paper IV. Besides, a major challenge highlighted in Papers II-III was the standardization of unstructured environmental inventory data into a digital dataset, complicated by municipality-based data silos, inconsistent documentation, hard copy formats, diverse file formats, and unclear inspection scopes. These issues likely stemmed from insufficient quality control of submitted inventories and lack of verification in post-demolition audits, resulting in a high volume of missing data and unreliable inspection records. Furthermore, the renewal of building registers posed limitations on accessing old inventories and hindered automated data matching.

Despite these challenges, the hazardous material dataset is unique and valuable, providing empirical records for identifying multiple hazardous materials, such as PCB-containing joints or sealants, double-glazed sealed windows, capacitors, acrylic flooring, and asbestos-containing materials, for instance pipe insulation, window or door insulation, cement panels, tiles or clinker, floor mat glue, floor mats, ventilation channels, and joints or sealants. High-quality, granular component data were primarily sourced from consultant reports or protocols detailing diagnostic results, as found in Conf I. By incorporating more comprehensive inventories from other municipalities, the study facilitated statistical analysis of inventoried records with wide geographical coverage, encompassing various building typologies and age cohorts.

Table 5.1 details the data constitution in each subgroup and highlights top-scoring hazardous materials across building classes and municipalities. Remarkably, the detection records of PCB and asbestos materials in large and complex buildings from Stockholm and Gothenburg municipalities showed promising potential. The detection rates for asbestos tile or clinker and pipe insulation in multifamily houses and school buildings were particularly noteworthy. To increase the training set size, data subgroups per building class were merged into residential and non-residential categories, with building categories and classes incorporated as features in predictive modeling.

Table 5.1. Score ranking of hazardous materials based on the data assessment matrix. Values in bold are the scores above 70 indicating data subgroups with high data quality and quantity, and hyphens imply missing values.

Hazardous material	Municipality	SF*	MF*	S*	C*	I*
ACM tile or clinker	Stockholm	22	**96**	**94**	**74**	**70**
	Gothenburg	34	**70**	**74**	**70**	**70**
	Malmö	0	0	50	25	25
	Kiruna	0	25	0	0	0
ACM pipe insulation	Stockholm	23	**93**	68	**74**	**71**
	Gothenburg	50	68	**74**	48	**70**
	Malmö	0	25	25	50	25
	Kiruna	0	68	0	0	0
ACM ventilation	Stockholm	0	**74**	48	**74**	**75**
	Gothenburg	0	44	**75**	46	50
	Malmö	-	0	25	25	0
	Kiruna	-	25	-	0	0
ACM floor mat	Stockholm	22	**70**	**71**	**74**	**74**
	Gothenburg	0	48	**75**	48	**74**
	Malmö	0	0	50	50	25
	Kiruna	0	0	-	0	0
ACM joint	Stockholm	24	**74**	69	**74**	**74**
	Gothenburg	0	50	50	24	50
	Malmö	-	0	50	50	25
	Kiruna	0	0	-	0	0
ACM floor mat glue	Stockholm	21	**71**	**70**	**74**	**70**
	Gothenburg	37	46	**74**	47	48
	Malmö	0	0	50	25	25
	Kiruna	0	0	0	0	0
ACM door or window insulation	Stockholm	21	**71**	**70**	**74**	**74**
	Gothenburg	36	47	**74**	48	48
	Malmö	-	-	0	25	25
	Kiruna	-	25	0	0	0
PCB joint	Stockholm	22	69	68	**71**	**70**
	Gothenburg	34	69	**74**	48	68
	Malmö	0	0	50	49	25
	Kiruna	0	-	0	0	-
PCB double-glazed sealed window	Stockholm	22	46	68	**71**	**71**
	Gothenburg	35	46	**74**	48	48
	Malmö	-	0	25	25	0
	Kiruna	0	0	0	0	0
ACM panel	Stockholm	19	48	68	**74**	**70**
	Gothenburg	32	46	50	44	46
	Malmö	-	0	25	25	25
	Kiruna	0	25	0	0	0

Hazardous material	Municipality	SF*	MF*	S*	C*	I*
PCB capacitor	Stockholm	24	38	68	46	68
	Gothenburg	52	44	**73**	46	66
	Malmö	-	0	25	0	0
	Kiruna	0	25	0	0	0
PCB acrylic floor	Stockholm	21	44	68	47	69
	Gothenburg	34	42	**74**	46	46
	Malmö	-	0	25	0	0
	Kiruna	-	-	0	-	-

*Building classes investigated were single-family houses (SF), multifamily houses (MF), school buildings (S), commercial or office buildings (C), and industrial buildings (I).

RQ2 *Which predictive methods can be used to estimate the presence of hazardous substances in buildings?*

Data-driven methods, such as statistical modeling, machine learning, and neural networks, can be utilized to predict the presence of hazardous substances in buildings and to identify their presence patterns to various extent. Among the twelve algorithms examined, tree-ensemble classifiers with added sample weights, including random forest, gradient boosting, XGBoost, and stacked ensemble models, demonstrated optimal performance. These classifiers were particularly effective in the binary classification of hazardous materials using input data from environmental inventories, as well as in the multi-class classification of indoor radon intervals based on measurement records. Table 5.2 presents the lead model types and their performance in predicting PCB and asbestos materials across different building categories, excerpted from Paper IV. In general, higher model performances were achieved in residential buildings. Asbestos door or window insulation and ventilation channels were predicted with high AUC scores above 0.9. Satisfactory prediction results (AUC: 0.65-0.88) were also obtained for PCB joints or sealants, capacitors, asbestos floor mat glue, cement panels, tile or clinker, pipe insulation, and joints or sealants in both residential and non-residential buildings.

Table 5.2. Performance of lead models for hazardous material prediction by building categories evaluated with AUC (1e-2) and listed in descending order.

Hazardous material	Residential	Non-residential	All buildings
ACM door insulation	Gradient Boosting 93	XGBoost 85	Gradient Boosting 84
ACM ventilation channel	Stacked ensemble 90	Random forest 74	Stacked ensemble 80
PCB joint	Gradient Boosting 88	XGBoost 75	Gradient Boosting 78
ACM floor mat glue	XGBoost/ Stacked ensemble 88	Gradient Boosting 73	Gradient Boosting 75
PCB capacitors	Generalized linear model 86	XGBoost 76	Random forest/ Gradient Boosting 82
ACM panel	Random forest 84	Random forest 65	Random forest/ Neural network 71
ACM tile/clinker	Gradient Boosting 83	Gradient Boosting 69	Gradient Boosting 75
ACM pipe insulation	Random forest 81	Gradient Boosting 75	Gradient Boosting 80
ACM joint/sealant	Random forest 78	Gradient Boosting 76	Random forest 78
PCB acrylic floor	-* -*	XGBoost 68	Stacked ensemble 66
ACM floor mat	XGBoost 61	Gradient Boosting 68	XGBoost 63
PCB double-glazed sealed window	XGBoost 61	Generalized linear model 66	XGBoost 60

*Detection of PCB acrylic floor in residential buildings was highly imbalanced and failed in training because of one cardinality (the number of possible values that a feature can assume).

In the comparative analysis in Paper VI, XGBoost models significantly outperformed neural network models in predicting indoor radon intervals across different building classes, achieving macro-F1 scores ranging from 0.93 to 0.96, compared to the neural network models' scores of 0.64 to 0.74. The annual average indoor radon concentrations from national measurements were categorized into three intervals, aligning with current legislative requirements. To address the imbalance in the indoor radon intervals dataset, where the low label (0-200 Bq/m³) was overrepresented and the medium labels (200-400 Bq/m³) and high labels (above 400 Bq/m³) were underrepresented, the study employed missing data imputation using *k*-NN (*k*-Nearest Neighbors) and resampling with SMOTE (Synthetic Minority Over-sampling Technique). Additionally, sample weight adjustments

were made in the algorithms. These methods resulted in an average error rate of 4.5% across building classes, except for single-family houses, which exhibited a slightly higher error rate of 6.5%. Notably, there were higher error rates in predicting the medium interval (6.7-10.8%) compared to the low (2.7-7.1%) and high intervals (1.5-2.7%). In contrast, regression modeling for indoor radon concentration prediction in Conf II was less successful, achieving lower performance than previous studies. This outcome suggests that while XGBoost models are highly effective in classification tasks for indoor radon prediction, traditional statistical and random forest regression modeling may face challenges in capturing the complexities of indoor radon concentration variations across different building types and conditions.

Table 5.3. Model performance for indoor radon prediction by building classes.

Algorithm	Single-family house	Multifamily house	School building	Other building
Multi-class classification with macro-F1 metric (1e-2)				
XGBoost	93	95	94	96
Neural network	66	74	64	68
Regression with R^2 metric (1e-2)				
Random forest	21	28	7	2
MARS	15	12	8	2

Learning Bayesian networks from environmental inventory data offers an alternative approach for unraveling the patterns of hazardous materials. This method investigated in Paper V involves structural and parameter learning to recognize factor dependencies, depicted in directed acyclic graphs (DAGs), and to compute conditional probability distributions. As illustrated for radioactive concrete in Table 5.4, various DAGs were generated using different types and numbers of input nodes, employing algorithms such as constraint-based estimator, max-min hill-climb, and tree search. The key advantage of Bayesian models lies in their superior output explainability, beneficial for post-modeling data analytics, as well as assisting causal inference. However, a trade-off exists in lower model resolution compared to machine learning or neural network models, as a result of numerical data binning used in approximate probability inference.

Additionally, the study explored the presence of radioactive concrete-containing materials across different building classes and investigated the impact of various building parameters, such as construction year, presence of basement, and ventilation types, on indoor radon concentrations. It was noted that the detection records for radioactive concrete components were statistically described and analyzed by building class, and predictive modeling was performed for estimating radioactive concrete in five Swedish municipalities. To enhance the models' performance and generalizability, including additional radioactive concrete inspection records from either environmental inventories or indoor radon measurements in more municipalities is necessary.

Table 5.4. Performance of Bayesian network models for radioactive concrete prediction.

Model	Algorithm	BIC	Edges in structure learning
1.1	Constraint-based estimator (pc)	-2729	P(Floor area, Building class, Basement, Radioactive concrete) = Pr (Building class | Radioactive concrete) Pr (Floor area | Building class) Pr (Floor area | Basement) Pr (Building class | Basement)
2.1	Max-min hill-climb (mmhc)	-3051	P(Construction year, Building class, Basement, Radioactive concrete) = Pr (Building class | Radioactive concrete) Pr (Basement | Construction year) Pr (Radioactive concrete | Construction year) Pr (Building class | Basement)
3.1	Max-min hill-climb (mmhc)/ tree search (ts)	-3123	P(Floor area, Average distance, Radioactive concrete) = Pr (Radioactive concrete | Distance to historical manufacturing plants) Pr (Distance to historical manufacturing plants | Floor area)

RQ3 *With what certainty can the presence of hazardous substances in the building stock be predicted?*

The predictive outcomes for the presence of hazardous substances, both for regional building stock and individual buildings, are provided in the form of probabilities and labels in Paper IV. The application scale for models of hazardous materials and radioactive concrete was set at the municipality building stock level, while indoor radon interval prediction models were applied to metropolitan building stock, based on geographical representativeness and the size of training samples. By utilizing the lead predictive models on building stock lacking environmental inventories or indoor radon measurements, the estimated shares of buildings or properties with a higher contamination probability were calculated, and their geospatial probability distribution visualized on building footprint maps. Model reliability and result certainty, particularly for indoor radon interval prediction, were evaluated by benchmarking the statistical shares against those predicted from existing measurements in Paper VI.

Prediction results suggest that the extent of contamination for the building stock built between 1930-1980 is considerably lower than the statistics from inspected and measured buildings. This discrepancy might be due to the overrepresentation of non-residential buildings (63%) and buildings constructed between the 1950s-1970s in the inventory data, along with more frequent measurements in buildings suspected of having radioactive concrete or high indoor radon exposure. In reality, and in prediction datasets, residential buildings constitute approximately 95% of the building stock in numbers, predominantly single-family houses, which are underrepresented in the inventory and indoor radon datasets, leading to higher uncertainty. Hence, the lower predicted contamination exposure

in the overall non-inventoried or unmeasured building stock aligns with current data, indicating a higher likelihood of detecting hazardous materials in medium to large postwar urban buildings. To refine the predictive models, it is necessary to increase the data size for underrepresented observations in the training set, based on the sample distribution of the prediction set, reflecting the building parameters of class, construction year, and building physical footprint. Radioactive concrete, conversely, was predicted to be more prevalent in the Swedish building stock than assumed in the BETSI survey (Swedish National Board of Housing Building and Planning (2010).

Overall, it appears that there is no single rule-of-thumb summarizing the presence patterns of hazardous substances; instead, these patterns should be contextualized within regional building stock and building typology. Critical features and their impact magnitude on the presence of in situ hazardous substances in the Swedish building stock were highlighted through model interpretation, employing feature importance heatmaps, SHAP values, and partial dependent plots. The presence of PCB and asbestos-containing materials was primarily associated with construction year, building physical footprint, floor area, and location (represented as postcode). Other specific building parameters related to certain hazardous materials also played a role, with patterns varying between building categories. According to the findings in Paper V, radioactive concrete was more commonly detected in multifamily houses, with its determinants being the distance to historical hazardous material manufacturing plants, building class, and construction year. Key attributes for predicting indoor radon included building physical footprint, floor area, construction year, coordinates, exhaust ventilation, uranium concentrations, basements, and natural ventilation, each with varying impact magnitudes across different building classes. These indicators not only confirmed some existing expert assumptions but also provided new insights, offering a holistic perspective on tracing in situ hazardous substances in the Swedish building stock.

5.2 Suggestions for Future Research

Future prospects based on the thesis work could be evolved in several directions:

Inventory dataset expansion on sampling size, geographical variety, and quantity information of hazardous materials

Currently, the developed models predict the probability of the occurrence of several asbestos and PCB components at the building level for municipalities with data samples, trained by building categories with limited granularity. If more observations from adjacent geographical areas become available, regional models could be created based on building class or construction period, considering region-

specific building practices. The assembly and compilation process would benefit from digital inventory protocols provided by municipalities. To enhance the models' generalizability to the national building stock, environmental inventories from all 286 Swedish municipalities are needed to obtain representative samples that reflect the diversity of regional building stock. The prototype models could then serve as checkpoint models, updating cost-efficiently with new observations in the training set. With expanding scope of data collection from environmental inventories, predictive modeling could advance beyond binary classification to quantify the amount of detected hazardous components. Such information, occasionally available in inventories, has the potential for descriptive analysis or regression modeling for estimating hazardous material quantities based on building typologies, volumes, location, and construction traditions. These environmental data could be reorganized into a cloud-based semantic data model, integrating them into geometric data such as BIM, building logbooks, and 3D point cloud models, ensuring data completeness and facilitating data retrieval for analysis and modeling.

Prototype model refinement and validation in real-world cases

The next critical step is refining and validating prototype models in real-world scenarios. Addressing sample selection bias and potential systematic bias in tree ensemble models through data preprocessing is vital. Control function approaches in machine learning regression (Brewer & Carlson, 2021) and empirical distribution matching (Belitz & Stackelberg 2021) could correct biases, though their effectiveness in classification models requires further research. Refining models through cost-sensitive learning for imbalanced classification, particularly weighting false negative errors of hazardous substances higher, could improve prediction performance. In addition, iterative experiments with different cost matrices could minimize misclassification errors. Validating the models in ongoing renovation or demolition projects is essential before integrating them into a web interface.

Continuous development of contamination risk assessment framework

The thesis has aimed to promote risk-informed pre-demolition audit inspections to design effective renovation and demolition planning strategies. Implementing and continuously developing the contamination risk assessment framework, comprising probability and consequence modules, is crucial. While the probability of hazardous material detection in existing buildings has been estimated, the extent of contamination requires further investigation, with inputs on the quantity of detected hazardous materials. Regarding the consequence of suspected hazardous materials, comprehensive information on practical field sampling and decontamination is needed. Quantifying these risk factors would enable the creation of an overarching decision support tool for estimating project time schedules, budgets, and working safety with regard to the presence of hazardous materials. The outcomes of the contamination risk assessment framework would include tailored renovation and demolition planning recommendations for efficient audits and waste management.

References

Abriha, D., Kovács, Z., Ninsawat, S., Bertalan, L., Balázs, B., & Szabó, S. (2018). Identification of roofing materials with discriminant function analysis and random forest classifiers on pan-sharpened worldview-2 imagery – A comparison. *Hungarian Geographical Bulletin*, *67*(4), 375–392. https://doi.org/10.15201/hungeobull.67.4.6

Adelikhah, M., Shahrokhi, A., Imani, M., Chalupnik, S., & Kovács, T. (2021). Radiological assessment of indoor radon and thoron concentrations and indoor radon map of dwellings in mashhad, Iran. *International Journal of Environmental Research and Public Health*, *18*(1), 1–15. https://doi.org/10.3390/ijerph18010141

Ajayi, S. O., Oyedele, L. O., Bilal, M., Akinade, O. O., Alaka, H. A., Owolabi, H. A., & Kadiri, K. O. (2015). Waste effectiveness of the construction industry: Understanding the impediments and requisites for improvements. *Resources, Conservation and Recycling*, *102*, 101–112. https://doi.org/10.1016/j.resconrec.2015.06.001

Akanbi, L. A., Oyedele, A. O., Oyedele, L. O., & Salami, R. O. (2020). Deep learning model for Demolition Waste Prediction in a circular economy. *Journal of Cleaner Production*, *274*. https://doi.org/10.1016/j.jclepro.2020.122843

Ankan, A., & Panda, A. (2015). pgmpy: Probabilistic Graphical Models using Python. *Proceedings of the 14th Python in Science Conference*, *Scipy*, 6–11. https://doi.org/10.25080/majora-7b98e3ed-001

Arevalillo, A., Hradil, P., Wahlström, M. (2017). *Technical and Economic Study with regard to the Development of Specific Tools and/or Guidelines for Assessment of Construction and Demolition Waste Streams prior to Demolition or Renovation of Buildings and Infrastructures*. *December*, 113.

Belitz, K., & Stackelberg, P. E. (2021). Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environmental Modelling and Software*, *139*(February), 105006. https://doi.org/10.1016/j.envsoft.2021.105006

Berggren, B., & Wall, M. (2019). Review of constructions and materials used in Swedish residential buildings during the post-war peak of production. *Buildings*, *9*(4). https://doi.org/10.3390/buildings9040099

Bergmans, J., Dierckx, P., & Broos, K. (2017). Semi-selective demolition : current demolition practices in Flanders. *HISER Conference*, *June*. https://doi.org/10.5281/zenodo.817324

Bergsdal, H., Brattebø, H., & Müller, D. B. (2014). Dynamic material flow analysis for PCBs in the Norwegian building stock. *Building Research and Information*, *42*(3), 359–370. https://doi.org/10.1080/09613218.2014.887898

Bodar, C., Spijker, J., Lijzen, J., Waaijers-van der Loop, S., Luit, R., Heugens, E., Janssen, M., Wassenaar, P., & Traas, T. (2018). Risk management of hazardous substances in a circular economy. *Journal of Environmental Management*, *212*, 108–114. https://doi.org/10.1016/j.jenvman.2018.02.014

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Bonifazi, G., Capobianco, G., & Serranti, S. (2018). Asbestos containing materials detection and classification by the use of hyperspectral imaging. *Journal of Hazardous Materials*, *344*, 981–993. https://doi.org/10.1016/j.jhazmat.2017.11.056

Bonifazi, G., Capobianco, G., & Serranti, S. (2019). Hyperspectral imaging and hierarchical PLS-DA applied to asbestos recognition in construction and demolition waste. *Applied Sciences (Switzerland)*, *9*(21), 1–15. https://doi.org/10.3390/app9214587

Brewer, D., & Carlson, A. (2021). *Addressing Sample Selection Bias for Machine Learning Methods*.

Byggföretagen. (2019). *Resource and waste guidelines for construction and demolition*. https://chalmers.instructure.com/courses/18307/files/2137211?wrap=1

Carbonari, A., Corneli, A., Di Giuda, G. M., Ridolfi, L., & Villa, V. (2019). A decision support system for multi-criteria assessment of large building stocks. *Journal of Civil Engineering and Management*, *25*(5), 477–494. https://doi.org/10.3846/jcem.2019.9872

Cerqueiro-Pequeño, J., Comesaña-Campos, A., Casal-Guisande, M., & Bouza-Rodríguez, J. B. (2021). Design and development of a new methodology based on expert systems applied to the prevention of indoor radon gas exposition risks. *International Journal of*

Environmental Research and Public Health, *18*(1), 1–33. https://doi.org/10.3390/ijerph18010269

Cha, G. W., Kim, Y. C., Moon, H. J., & Hong, W. H. (2017). New approach for forecasting demolition waste generation using chi-squared automatic interaction detection (CHAID) method. *Journal of Cleaner Production*, *168*, 375–385. https://doi.org/10.1016/j.jclepro.2017.09.025

Cha, G. W., Moon, H. J., Kim, Y. M., Hong, W. H., Hwang, J. H., Park, W. J., & Kim, Y. C. (2020). Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets. *International Journal of Environmental Research and Public Health*, *17*(19), 1–15. https://doi.org/10.3390/ijerph17196997

Chen, C. (2017). Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science*, *2*(2), 1–40. https://doi.org/10.1515/JDIS-2017-0006

Chen, S., Fang, K., Dhakalc, S., Kharrazi, A., Tong, K., & Ramaswami, A. (2023). Advancing urban infrastructure research for a carbon-neutral and sustainable future. *Resources , Conservation & Recycling*, *197*, 2020–2023. https://doi.org/10.1016/j.resconrec.2023.107049

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, *137*, 43–90. www.elsevier.com/locate/artint

Clavensjö, B., & Åkerblom, G. (2020). *Radonboken. Befintliga byggnader* (Fjärde utg). Svensk byggtjänst.

Clemmensen, L. H., & Kjærsgaard, R. D. (2022). Data Representativity for Machine Learning and AI Systems. *FAccT '22: ACM Conference on Fairness, Accountability and Transparency, June 21-24, 2022, Seoul, South Korea*, *1*(1), 1–16. http://arxiv.org/abs/2203.04706

Cook, E., Velis, C. A., & Black, L. (2022). Construction and Demolition Waste Management: A Systematic Scoping Review of Risks to Occupational and Public Health. *Frontiers in Sustainability*, *3*(June). https://doi.org/10.3389/frsus.2022.924926

Copes, R., & Peterson, E. (2014). *Indoor Radon a Public Health Perspective*.

Csiszar, S. A., Diamond, M. L., & Daggupaty, S. M. (2014). The magnitude and spatial range of current-use urban PCB and PBDE emissions estimated using a coupled multimedia and air transport model. *Environmental Science & Technology*, *48*(2), 1075–1083. https://doi.org/10.1021/ES403080T

Darko, A., Chan, A. P. C., Adabre, M. A., Edwards, D. J., Hosseini, M. R., & Ameyaw, E. E. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities.

Automation in Construction, *112*(January). https://doi.org/10.1016/j.autcon.2020.103081

Deepika, R., Markopoulou, A., Marengo, M., & Wolf, C. de. (2022). Enabling component reuse from existing buildings through machine learning. *Conference: Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*, *2*, 577–586.

Diamond, M. L., Melymuk, L., Csiszar, S. A., & Robson, M. (2010). Estimation of PCB stocks, emissions, and urban fate: Will our policies reduce concentrations and exposure? *Environmental Science and Technology*, *44*(8), 2777–2783. https://doi.org/10.1021/es9012036

Diefenbacher, P. S., Bogdal, C., Gerecke, A. C., Glüge, J., Schmid, P., Scheringer, M., & Hungerbühler, K. (2015). Emissions of polychlorinated biphenyls in Switzerland: A combination of long-term measurements and modeling. *Environmental Science and Technology*, *49*(4), 2199–2206. https://doi.org/10.1021/ES505242D/SUPPL_FILE/ES505242D_SI_001.PDF

Diefenbacher, P. S., Gerecke, A. C., Bogdal, C., & Hungerbühler, K. (2016). Spatial Distribution of Atmospheric PCBs in Zurich, Switzerland: Do Joint Sealants Still Matter? *Environmental Science and Technology*, *50*(1), 232–239. https://doi.org/10.1021/acs.est.5b04626

Donovan, S., & Pickin, J. (2016a). An Australian stocks and flows model for asbestos. *Waste Management and Research*, *34*(10), 1081–1088. https://doi.org/10.1177/0734242X16659353

Donovan, S., & Pickin, J. (2016b). An Australian stocks and flows model for asbestos. *Waste Management and Research*, *34*(10), 1081–1088. https://doi.org/10.1177/0734242X16659353

Döringer, S. (2021). 'The problem-centred expert interview'. Combining qualitative interviewing approaches for investigating implicit expert knowledge. *International Journal of Social Research Methodology*, *24*(3), 265–278. https://doi.org/10.1080/13645579.2020.1766777

ECORYS. (2016). *EU Construction & Demolition Waste Management Protocol*.

Elío, J., Cinelli, G., Bossew, P., Gutiérrez-Villanueva, J. L., Tollefsen, T., De Cort, M., Nogarotto, A., & Braga, R. (2019). The first version of the Pan-European Indoor Radon Map. *Natural Hazards and Earth System Sciences*, *19*(11), 2451–2464. https://doi.org/10.5194/nhess-19-2451-2019

European Chemicals Agency. (2007). *REACH legislation*. https://echa.europa.eu/regulations/reach/understanding-reach

European Commission. (2008). *Waste Framework Directive*. https://ec.europa.eu/environment/topics/waste-and-recycling/waste-framework-directive_en

European Commission. (2018). 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*, *4724185*, 37.

European Commission. (2019). *A European Green Deal*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

European Commission. (2020a). *A new Circular Economy Action Plan*. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1583933814386&uri=COM:2020:98:FIN

European Commission. (2020b). *A Renovation Wave for Europe*. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1603122220757&uri=CELEX:52020DC0662

European Commission. (2021). *COMMISSION STAFF WORKING DOCUMENT: Scenarios for a transition pathway for a resilient, greener and more digital construction ecosystem*.

European Commission. (2022). *Policy brief Level(s) and the New European Bauhaus*.

Franzblau, A., Demond, A. H., Sayler, S. K., D'Arcy, H., & Neitzel, R. L. (2020). Asbestos-containing materials in abandoned residential dwellings in Detroit. *Science of the Total Environment*, *714*, 136580. https://doi.org/10.1016/j.scitotenv.2020.136580

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics*.

Garfield, E. (2004). Historiographic Mapping of Knowledge Domains Literature. *Journal of Information Science*, *30*(2), 119–145. https://doi.org/10.1177/0165551504042802

Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, *8*(2), 209–218. https://doi.org/10.1142/S0129065797000227

Gibril, M. B. A., Shafri, H. Z. M., & Hamedianfar, A. (2017). New semi-automated mapping of asbestos cement roofs using rule-based object-based image analysis and Taguchi optimization technique from WorldView-2 images. *International Journal of Remote Sensing*, *38*(2), 467–491. https://doi.org/10.1080/01431161.2016.1266109

Glüge, J., Steinlin, C., Schalles, S., Wegmann, L., Tremp, J., Breivik, K., Hungerbühler, K., & Bogdal, C. (2017). Import, use, and emissions of

PCBs in Switzerland from 1930 to 2100. In *PLoS ONE* (Vol. 12, Issue 10). https://doi.org/10.1371/journal.pone.0183768

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Govorko, M., Fritschi, L., & Reid, A. (2018). Accuracy of a mobile app to identify suspect asbestos-containing material in Australian residential settings. *Journal of Occupational and Environmental Hygiene*, *15*(8), 598–606. https://doi.org/10.1080/15459624.2018.1475743

Govorko, M., Fritschi, L., & Reid, A. (2019). Using a mobile phone app to identify and assess remaining stocks of in situ asbestos in Australian residential settings. *International Journal of Environmental Research and Public Health*, *16*(24). https://doi.org/10.3390/ijerph16244922

Govorko, M., Fritschi, L., White, J., & Reid, A. (2017). Identifying Asbestos-Containing Materials in Homes: Design and Development of the ACM Check Mobile Phone App. *JMIR Formative Research*, *1*(1), e7. https://doi.org/10.2196/formative.8370

Grönberg, A. L. (2017). *Materialinventering av byggnader inför rivning*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature 2020 585:7825*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (Issue 2). Springer. https://doi.org/10.1198/jasa.2004.s339

Hong, T., Wang, Z., Luo, X., & Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, *212*, 109831. https://doi.org/10.1016/j.enbuild.2020.109831

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.

Johansson, T., Olofsson, T., & Mangold, M. (2017). Development of an energy atlas for renovation of the multifamily building stock in Sweden. *Applied Energy*, *203*, 723–736. https://doi.org/10.1016/j.apenergy.2017.06.027

Kartal Koc, E., & Bozdogan, H. (2015). Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function. *Machine Learning*, *101*(1–3), 35–58. https://doi.org/10.1007/s10994-014-5440-5

Khan, S. M., Pearson, D. D., Rönnqvist, T., Nielsen, M. E., Taron, J. M., & Goodarzi, A. A. (2021). Rising Canadian and falling Swedish radon gas exposure as a consequence of 20th to 21st century residential build practices. *Scientific Reports*, *11*(1), 1–15. https://doi.org/10.1038/s41598-021-96928-x

Kim, H. J., Hamann, R., Sotiralis, P., Ventikos, N. P., & Straub, D. (2018). Bayesian network for risk-informed inspection planning in ships. *Beton-Und Stahlbetonbau*, *113*(2), 116–121. https://doi.org/10.1002/best.201800054

Kim, J. T., & Yu, C. W. F. (2014). Hazardous materials in buildings. *Indoor and Built Environment*, *23*(1), 44–61. https://doi.org/10.1177/1420326X14524073

Kim, Y. C., & Hong, W. H. (2017). Optimal management program for asbestos containing building materials to be available in the event of a disaster. *Waste Management*, *64*, 272–285. https://doi.org/10.1016/j.wasman.2017.03.042

Kohler, N. (2018). From the design of green buildings to resilience management of building stocks. *Building Research and Information*, *46*(5), 578–593. https://doi.org/10.1080/09613218.2017.1356122

Kolarik, B., Frederiksen, M., Meyer, H. W., Ebbehøj, N. E., & Gunnarsen, L. B. (2016). Investigation of the importance of tertiary contamination, temperature and human behaviour on PCB concentrations in indoor air. *Indoor and Built Environment*, *25*(1), 229–241. https://doi.org/10.1177/1420326X14543505

Koutamanis, A., van Reijn, B., & van Bueren, E. (2018). Urban mining and buildings: A review of possibilities and limitations. *Resources, Conservation and Recycling*, *138*(July), 32–39. https://doi.org/10.1016/j.resconrec.2018.06.024

Kropat, G., Bochud, F., Jaboyedoff, M., Laedermann, J. P., Murith, C., Palacios Gruson, M., & Baechler, S. (2015). Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: An application to Switzerland. *Science of the Total Environment*, *505*, 137–148. https://doi.org/10.1016/j.scitotenv.2014.09.064

Kropat, G., Bochud, F., Jaboyedoff, M., Laedermann, J. P., Murith, C., Palacios, M., & Baechler, S. (2014). Major influencing factors of indoor radon concentrations in Switzerland. *Journal of Environmental*

Radioactivity, *129*, 7–22. https://doi.org/10.1016/J.JENVRAD.2013.11.010

Kropat, G., Bochud, F., Jaboyedoff, M., Laedermann, J. P., Murith, C., Palacios, M., & Baechler, S. (2015). Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units. *Journal of Environmental Radioactivity*, *147*, 51–62. https://doi.org/10.1016/j.jenvrad.2015.05.006

Krówczyńska, M., Raczko, E., Staniszewska, N., & Wilk, E. (2020). Asbestos-cement roofing identification using remote sensing and convolutional neural networks (CNNs). *Remote Sensing*, *12*(3), 1–16. https://doi.org/10.3390/rs12030408

Lain, S. (2017). Show, don't tell: Reading workshop fosters engagement and success. *Texas Journal of Literacy Education*, *5*(2).

Lewis, M. (2019). Incompatible trends - Hazardous Chemical Usage in Building Products Poses Challenges for Functional Circular Construction. *IOP Conference Series: Earth and Environmental Science*, *225*(1), 012046. https://doi.org/10.1088/1755-1315/225/1/012046

López Ruiz, L. A., Roca Ramón, X., & Gassó Domingo, S. (2020). The circular economy in the construction and demolition waste sector – A review and an integrative model approach. *Journal of Cleaner Production*, *248*. https://doi.org/10.1016/j.jclepro.2019.119238

Lucchi, E., Exner, D., & D'Alonzo, V. (2018). Building stock analysis as a method to assess the heritage value and the energy performance of an Alpine historical urban settlement. *Energy Efficiency in Historic Buildings 2018*, *53*(9), 482–492.

Mangold, M., Österbring, M., & Wallbaum, H. (2015). Handling data uncertainties when using Swedish energy performance certificate data to describe energy usage in the building stock. *Energy and Buildings*, *102*, 328–336. https://doi.org/10.1016/j.enbuild.2015.05.045

McCullagh, P., & Nelder, JA. (1983). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

Mecharnia, T., Khelifa, L. C., Pernelle, N., & Hamdi, F. (2019). An approach toward a prediction of the presence of asbestos in buildings based on incomplete temporal descriptions of marketed products. *K-CAP 2019 - Proceedings of the 10th International Conference on Knowledge Capture*, 239–242. https://doi.org/10.1145/3360901.3364428

Melymuk, L., Robson, M., Helm, P. A., & Diamond, M. L. (2013). Application of land use regression to identify sources and assess spatial

variation in urban SVOC concentrations. *Environmental Science & Technology*, *47*(4), 1887–1895. https://doi.org/10.1021/ES3043609

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7). https://doi.org/10.1371/journal.pmed.1000097

Nam, I. S., Oh, H. J., Kim, J. M., Yang, J. H., Kim, J. S., & Sohn, J. R. (2015). Comparison of risk assessment criteria and distribution of asbestos-containing materials in school building. *International Journal of Environmental Research*, *9*(4), 1341–1350. https://doi.org/10.22059/ijer.2015.1026

Norvig, P., & Russell, S. (2021). *Artificial Intelligence: A Modern Approach, 4th US ed.* Pearson. http://aima.cs.berkeley.edu/

Nußholz, J., Çetin, S., Eberhardt, L., De Wolf, C., & Bocken, N. (2023). From circular strategies to actions: 65 European circular building cases and their decarbonisation potential. *Resources, Conservation and Recycling Advances*, *17*(January). https://doi.org/10.1016/j.rcradv.2023.200130

Olsthoorn, B., Rönnqvist, T., Lau, C., Rajasekaran, S., Persson, T., Månsson, M., & Balatsky, A. V. (2022). Indoor radon exposure and its correlation with the radiometric map of uranium in Sweden. *Science of the Total Environment*, *811*. https://doi.org/10.1016/j.scitotenv.2021.151406

Oni, O. M., Aremu, A. A., Oladapo, O. O., Agboluaje, B. A., & Fajemiroye, J. A. (2022). Artificial neural network modeling of meteorological and geological influences on indoor radon concentration in selected tertiary institutions in Southwestern Nigeria. *Journal of Environmental Radioactivity*, *251–252*(June), 106933. https://doi.org/10.1016/j.jenvrad.2022.106933

Parada, H., Sun, X., Tse, C. K., Engel, L. S., Hoh, E., Olshan, A. F., & Troester, M. A. (2020). Plasma levels of polychlorinated biphenyls (PCBs) and breast cancer mortality: The Carolina Breast Cancer Study. *International Journal of Hygiene and Environmental Health*, *227*, 113522. https://doi.org/10.1016/J.IJHEH.2020.113522

Pasichnyi, O., Wallin, J., Levihn, F., Shahrokni, H., & Kordas, O. (2019). Energy performance certificates — New opportunities for data-enabled urban energy policy instruments? *Energy Policy*, *127*(October 2018), 486–499. https://doi.org/10.1016/j.enpol.2018.11.051

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pereira, C., De Brito, J., & Silvestre, J. D. (2020). Global inspection, diagnosis and repair system for buildings: Homogenising the classification of diagnosis methods. *Rehabend*, *June*, 554–562.

Pereira, C., Silva, A., Ferreira, C., de Brito, J., Flores-Colen, I., & Silvestre, J. D. (2021). Uncertainty in building inspection and diagnosis: A probabilistic model quantification. *Infrastructures*, *6*(9). https://doi.org/10.3390/infrastructures6090124

Powell, J., Jain, P., Bigger, A., & Townsend, T. G. (2015). 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. *Journal of Hazardous, Toxic, and Radioactive Waste*, *19*(4), 05015001. https://doi.org/10.1061/(asce)hz.2153-5515.0000266

Raczko, E., Krówczyńska, M., & Wilk, E. (2022). Asbestos roofing recognition by use of convolutional neural networks and high-resolution aerial imagery. Testing different scenarios. *Building and Environment*, *217*(February). https://doi.org/10.1016/j.buildenv.2022.109092

Raschka, S., & Mirjalili, V. (2019). *Python machine learning : machine learning and deep learning with python, scikit-learn, and tensorflow 2 - 3rd Edition*. Packt Publishing Ltd.

Rašković, M., Ragossnig, A. M., Kondracki, K., & Ragossnig-Angst, M. (2020). Clean construction and demolition waste material cycles through optimised pre-demolition waste audit documentation: A review on building material assessment tools. *Waste Management and Research*, *38*(9), 923–941. https://doi.org/10.1177/0734242X20936763

Robson, M., Melymuk, L., Csiszar, S. A., Giang, A., Diamond, M. L., & Helm, P. A. (2010a). Continuing sources of PCBs: The significance of building sealants. *Environment International*, *36*(6), 506–513. https://doi.org/10.1016/j.envint.2010.03.009

Robson, M., Melymuk, L., Csiszar, S. A., Giang, A., Diamond, M. L., & Helm, P. A. (2010b). Continuing sources of PCBs: The significance of building sealants. *Environment International*, *36*(6), 506–513. https://doi.org/10.1016/j.envint.2010.03.009

Rönnqvist, T. (2021). *Analysis of Radon Levels in Swedish Dwellings and Workplaces*.

Ruder, A. M., Hein, M. J., Hopf, N. B., & Waters, M. A. (2014). Mortality among 24,865 workers exposed to polychlorinated biphenyls (PCBs) in three electrical capacitor manufacturing plants: A ten-year update. *International Journal of Hygiene and Environmental Health*, *217*(0), 176. https://doi.org/10.1016/J.IJHEH.2013.04.006

Rudy, J. (2013). *py-earth documentation*. https://contrib.scikit-learn.org/py-earth/content.html

Sandberg, S., & Hultegård, L. (2021). *Cirkulära produktflöden i byggsektorn för ökad resurseffektivitet*. Linköping University.

Sarra, A., Fontanella, L., Valentini, P., & Palermi, S. (2016). Quantile regression and Bayesian cluster detection to identify radon prone areas. *Journal of Environmental Radioactivity*, *164*, 354–364. https://doi.org/10.1016/j.jenvrad.2016.06.014

Schat, E., van de Schoot, R., Kouw, W. M., Veen, D., & Mendrik, A. M. (2020). The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. *PLoS ONE*, *15*(8 August), 1–16. https://doi.org/10.1371/journal.pone.0237009

Sedin, D., & Hjelte, I. (2004). *The Radon Situation in Sweden* (Issue July).

Shanahan, C. E., Spak, S. N., Martinez, A., & Hornbuckle, K. C. (2015). Inventory of PCBs in Chicago and Opportunities for Reduction in Airborne Emissions and Human Exposure. *Environmental Science and Technology*, *49*(23), 13878–13888. https://doi.org/10.1021/acs.est.5b00906

Shi, W., Goodchild, M., Batty, M., Kwan, M.-P., & Zhang, A. (2021). Urban Informatics. In *Urban Book Series*. https://doi.org/10.1007/978-981-15-8983-6_16

Song, S.-J., Jang, B.-K., Jo, B.-H., Kim, Y.-J., Heo, E.-H., Lee, J.-D., Son, B.-S., & Lee, J.-W. (2016). An Asbestos Risk Assessment and Areal Distribution of Asbestos Containing Materials in Public Buildings. *Journal of Korean Society of Occupational and Environmental Hygiene*, *26*(3), 267–276. https://doi.org/10.15269/jksoeh.2016.26.3.267

STHDA. (2017). *MCA - Multiple Correspondence Analysis in R*. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/#dimension-description

Swedish Land Survey. (2022). *Överföringsformatet i Fastighetsregistret*. 1–263.

Swedish Land Survey, Swedish Maritime Administration, Geological Survey of Sweden, & Sweden, S. (2022). *Geodata Extraction Tool*. https://zeus.slu.se/get/?drop=

Swedish National Board of Housing Building and Planning. (2010). *Technical status in Swedish buildings - results from the BETSI project (Teknisk status i den svenska bebyggelsen - resultat från projektet BETSI)*. http://www.boverket.se/globalassets/publikationer/dokument/2011/betst-teknisk-status.pdf

Swedish National Board of Housing Building and Planning. (2023). *Boverket ska hjälpa byggsektorn att utvecklas mot en cirkulär ekonomi*. https://www.regeringen.se/pressmeddelanden/2022/02/boverket-ska-hjalpa-byggsektorn-att-utvecklas-mot-en-cirkular-ekonomi/

UN Habitat. (2022). *World Cities Report*. https://unhabitat.org/wcr/

University of Naples Federico II. (2023). *Bibliometrix R Package*. https://www.bibliometrix.org/

Valcarce, D., Alvarellos, A., Rabuñal, J. R., Dorado, J., & Gestal, M. (2022). Machine Learning-Based Radon Monitoring System. *Chemosensors*, *10*(7). https://doi.org/10.3390/chemosensors10070239

Villoria Sáez, P., & Osmani, M. (2019). A diagnosis of construction and demolition waste generation and recovery practice in the European Union. *Journal of Cleaner Production*, *241*. https://doi.org/10.1016/j.jclepro.2019.118400

Wahlström, M., Bergmans, J., Teittinen, T., Bachér, J., Smeets, A., & Paduart, A. (2020). Construction and Demolition Waste : challenges and opportunities in a circular economy. In *Eionet Report - ETC/WMGE 2020/1.* (Issue January). https://www.eea.europa.eu/publications/construction-and-demolition-waste-challenges/at_download/file

Wahlström, M., Teittinen, T., Kaartinen, T., & Liesbet, van C. (2019). *Hazardous substances in construction products and materials: PARADE. Best practices for Pre-demolition Audits ensuring high quality RAw materials*.

Wahlström, M., Zu Castell-Rüdenhausen, M., Hradil, P., Smith, K. H., Oberender, A., Ahlm, M., Götbring, J., & Hansen, J. B. (2019). *Improving quality of construction & demolition waste-Requirements for pre-demolition audit*. https://doi.org/10.6027/TN2019-508

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

Węglarz, A., & Gilewski, P. G. (2017). Application of expert systems in the construction sector. *MATEC Web of Conferences*, *117*, 4–10. https://doi.org/10.1051/matecconf/201711700176

Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J. C., & Mandin, C. (2019). Machine learning and statistical models for predicting indoor air quality. *Indoor Air*, *29*(5), 704–726. https://doi.org/10.1111/ina.12580

Westerholm, P., Remaéus, B., & Svartengren, M. (2017). The tale of asbestos in Sweden 1972-1986— the pathway to a near-total ban. *International Journal of Environmental Research and Public Health*, *14*(11), 1–11. https://doi.org/10.3390/ijerph14111433

Wilk, E., Krówczyńska, M., & Pabjanek, P. (2015). Determinants influencing the amount of asbestos-cement roofing in Poland. *Miscellanea Geographica*, *19*(3), 82–86. https://doi.org/10.1515/mgrsd-2015-0014

Wilk, E., Krówczyńska, M., Pabjanek, P., & Mędrzycki, P. (2017). Estimation of the amount of asbestos-cement roofing in Poland. *Waste Management and Research*, *35*(5), 491–499. https://doi.org/10.1177/0734242X16683271

Wilk, E., Krówczyńska, M., & Zagajewski, B. (2019). Modelling the spatial distribution of asbestos-cement products in Poland with the use of the random forest algorithm. *Sustainability (Switzerland)*, *11*(16). https://doi.org/10.3390/su11164355

Wu, H., Zuo, J., Zillante, G., Wang, J., & Yuan, H. (2019). Status quo and future directions of construction and demolition waste research: A critical review. *Journal of Cleaner Production*, *240*, 118163. https://doi.org/10.1016/j.jclepro.2019.118163

Xiang, L., Tan, Y., Shen, G., & Jin, X. (2022). Applications of multi-agent systems from the perspective of construction management: A literature review. *Engineering, Construction and Architectural Management*, *29*(9), 3288–3310. https://doi.org/10.1108/ECAM-01-2021-0038

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction*, *119*(August 2019), 103331. https://doi.org/10.1016/j.autcon.2020.103331

Yang, Z., Xue, F., & Lu, W. (2021). Handling missing data for construction waste management: machine learning based on aggregated waste generation behaviors. *Resources, Conservation and Recycling*, *175*(April), 105809. https://doi.org/10.1016/j.resconrec.2021.105809

Yu, B., Wang, J., Li, J., Zhang, J., Lai, Y., & Xu, X. (2019). Prediction of large-scale demolition waste generation during urban renewal: A hybrid

trilogy method. *Waste Management*, *89*, 1–9. https://doi.org/10.1016/j.wasman.2019.03.063

Zoraja, B., Ubavin, D., Stanisavljevic, N., Vujovic, S., Mucenski, V., Hadzistevic, M., & Bjelica, M. (2021). Assessment of asbestos and asbestos waste quantity in the built environment of transition country. *Waste Management and Research*. https://doi.org/10.1177/0734242X211064031

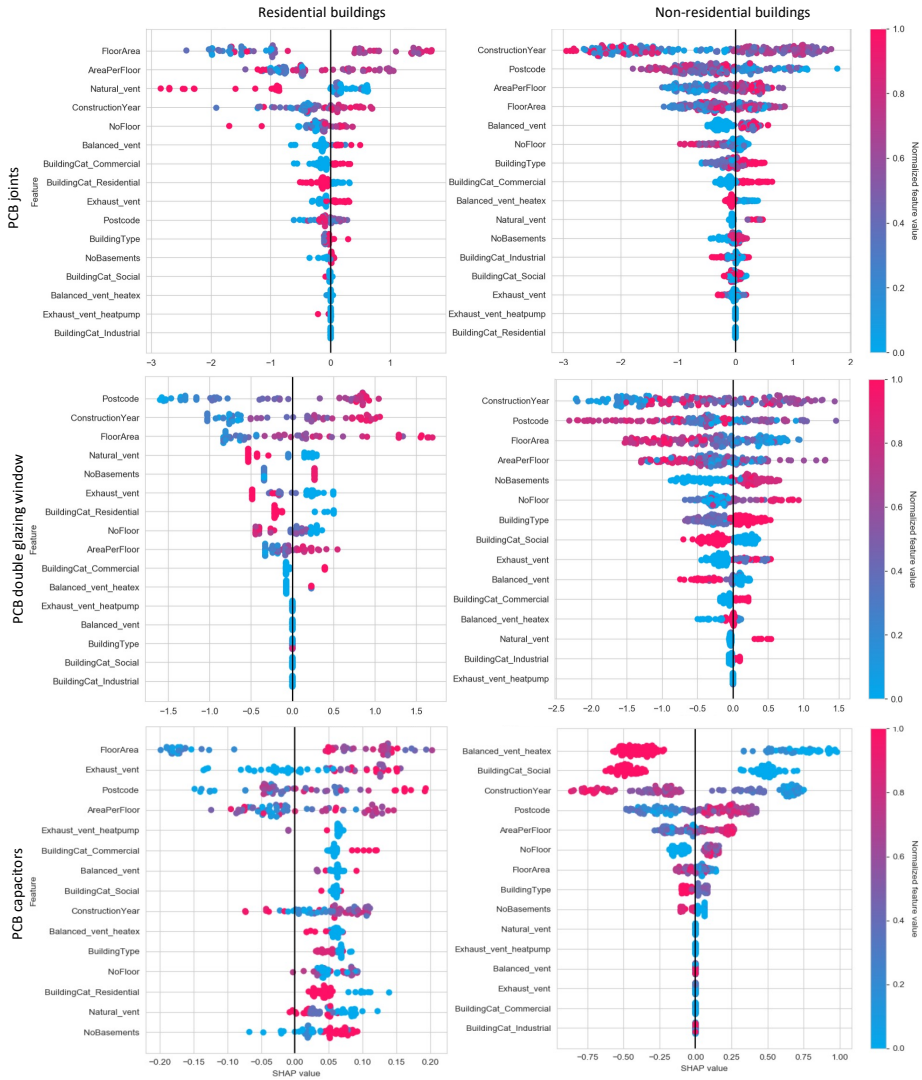
Appendix I Literature on Building Contamination Applications

Table A1. Literature on data-driven building contamination prediction.

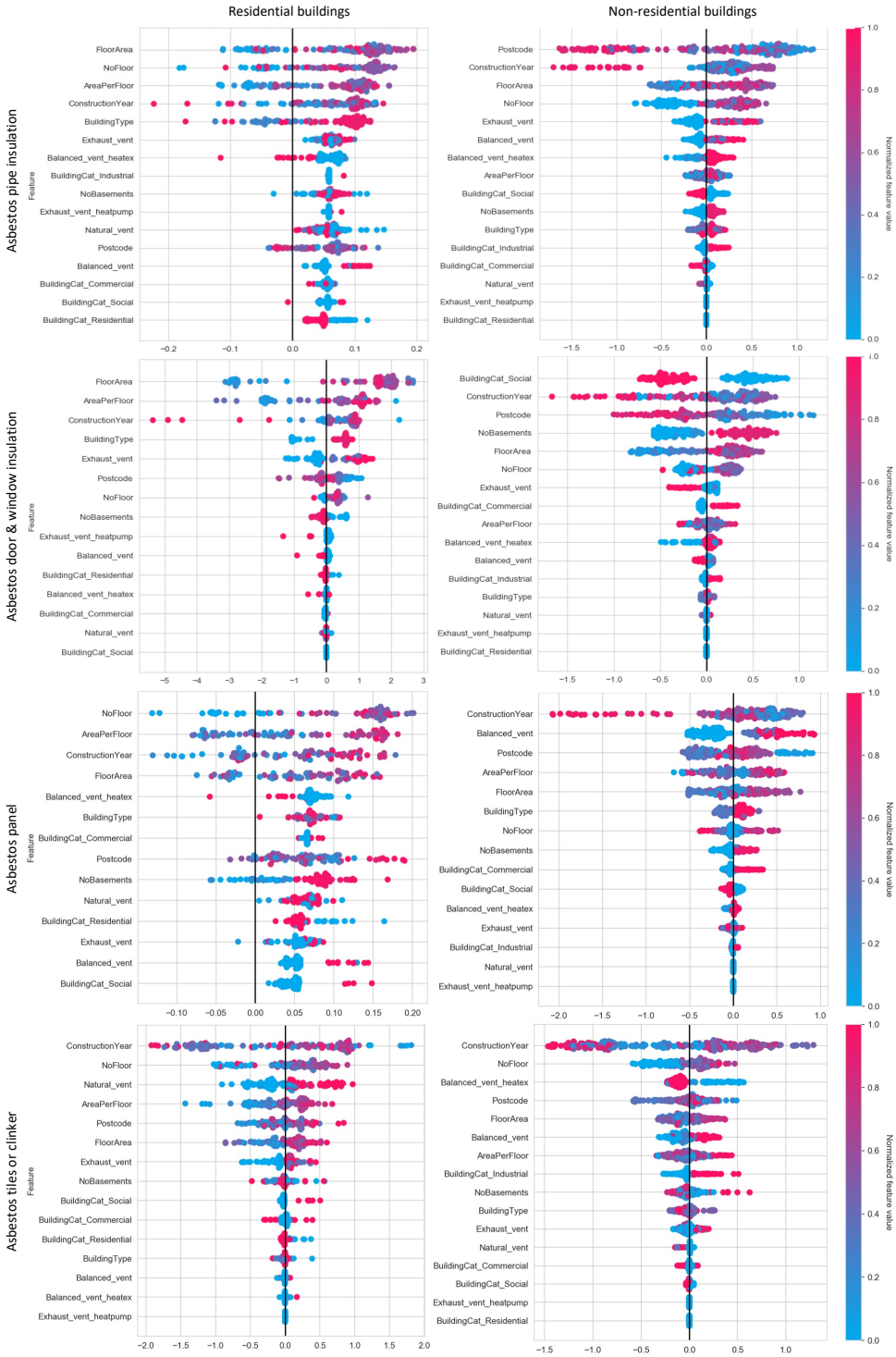
Substance	Application	Method	Reference
Asbestos	Asbestos building stock quantification	Statistical modeling (descriptive analysis; material flow analysis)	(Donovan & Pickin, 2016b; Zoraja et al., 2021)
	ACM characterization	Statistical modeling (descriptive analysis)	(Franzblau et al., 2020)
	ACM separation	Statistical modeling (Sampling)	(Powell et al., 2015)
	ACM identification	Statistical modeling (descriptive analysis; mobile app)	(Govorko et al., 2017, 2018, 2019)
	ACM prediction	Statistical modeling (ontology and rule-based methods)	(Mecharnia et al., 2019)
	Feature identification for ACM roofing	Statistical modeling (correlation)	(Wilk et al., 2015)
	ACM roofing amount identification and estimation	Machine learning (random forest)	(Wilk et al., 2017, 2019)
		Deep learning (remote sensing; CNN)	(Krówczyńska et al., 2020; Raczko et al., 2022)
PCB	PCB building stock quantification	Statistical modeling (descriptive analysis; material flow analysis; dynamic mass flow)	(Bergsdal et al., 2014; Glüge et al., 2017)
	PCB emission estimation	Statistical modeling (mass-balance fugacity model; GIS)	(Diamond et al., 2010; Robson et al., 2010b)
		Machine learning (field sampling; GIS; PCA)	(Kolarik et al., 2016)
		Statistical modeling (Concentration propagation)	(Shanahan et al., 2015)

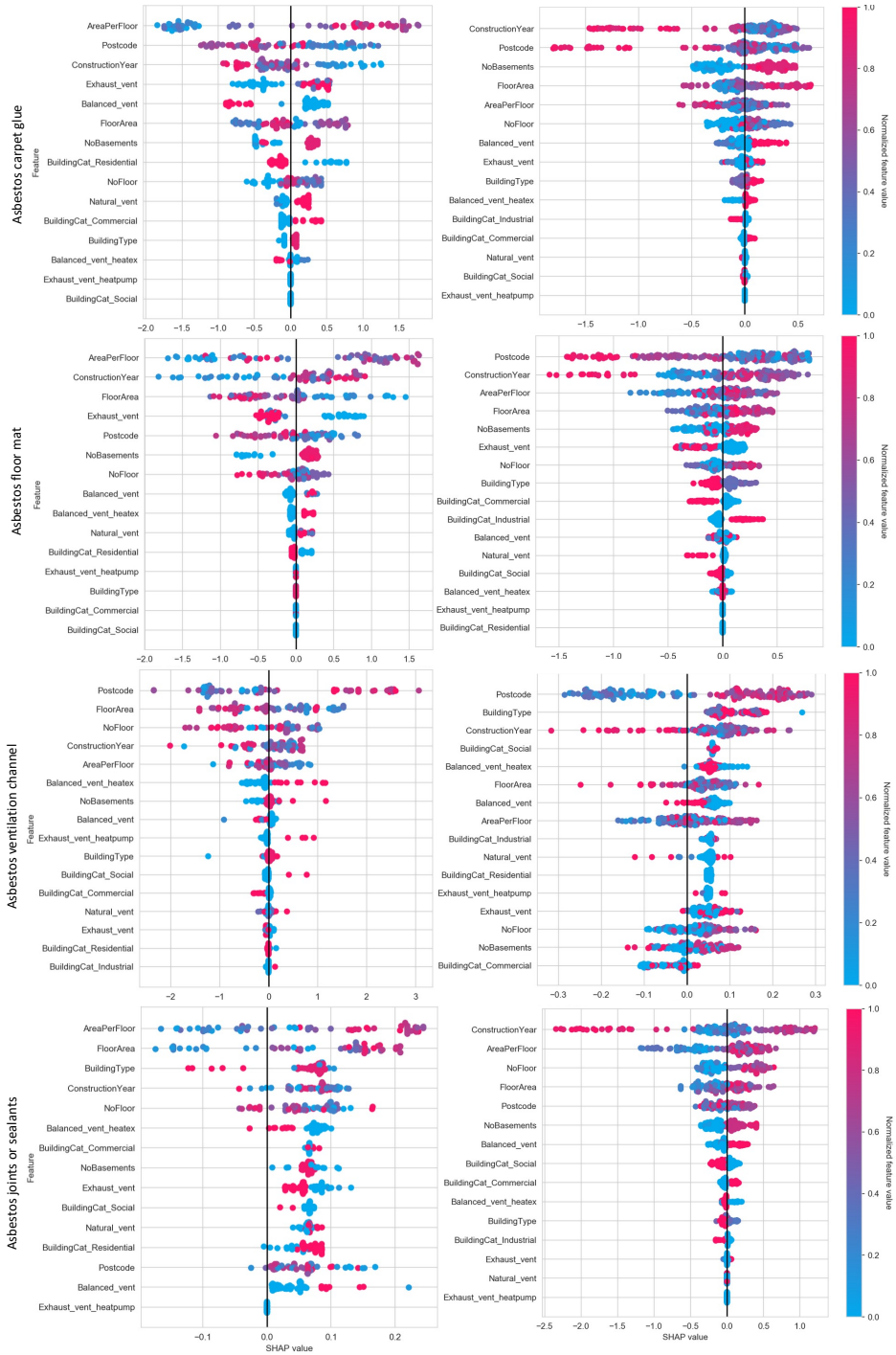
Substance	Application	Method	Reference
PCB	Spatial distribution of PCB estimation	Statistical modeling (Gaussian diffusion model; multi-compartment box model)	(Diefenbacher et al., 2015, 2016)
		Statistical modeling (regression model)	(Melymuk et al., 2013)
		Statistical modeling (multimedia and air transport model)	(Csiszar et al., 2014)
Indoor radon	Continental or national indoor radon map generation	Statistical modeling (spatial inference; interpolation)	(Adelikhah et al., 2021; Elío et al., 2019)
	National indoor radon prediction	Statistical modeling (kernel regression; probability estimation)	(Kropat et al., 2015a; Kropat et al., 2014)
		Machine learning (random forest; Bayesian additive regression trees; k-medoids clustering)	(Kropat et al., 2015b)
		Deep learning (ANN)	(Oni et al., 2022)
	Spatial clusters of radon-prone areas	Statistical modeling (Bayesian spatial quantile regression; stepwise analysis)	(Sarra et al., 2016)
	Indoor radon monitoring	Deep learning (RNN)	(Khan et al., 2021; Valcarce et al., 2022)

Appendix II SHAP Summary Plots for Hazardous Materials Prediction



(i) PCB-containing materials

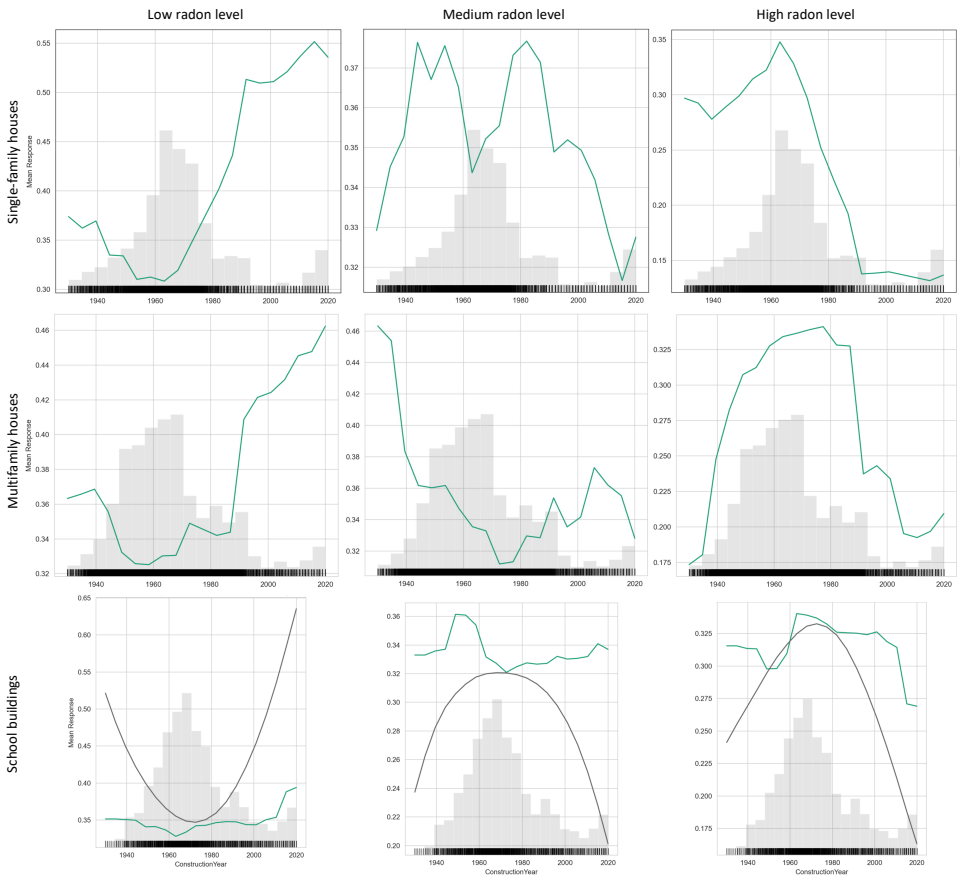

Residential buildings
Non-residential buildings
Asbestos pipe insulation
Asbestos door & window insulation
Asbestos panel
Asbestos tiles or clinker
Feature
Normalized feature value
FloorArea
NoFloor
AreaPerFloor
ConstructionYear
BuildingType
Exhaust_vent
Balanced_vent_heatex
BuildingCat_Industrial
NoBasements
Exhaust_vent_heatpump
Natural_vent
Postcode
Balanced_vent
BuildingCat_Commercial
BuildingCat_Social
BuildingCat_Residential



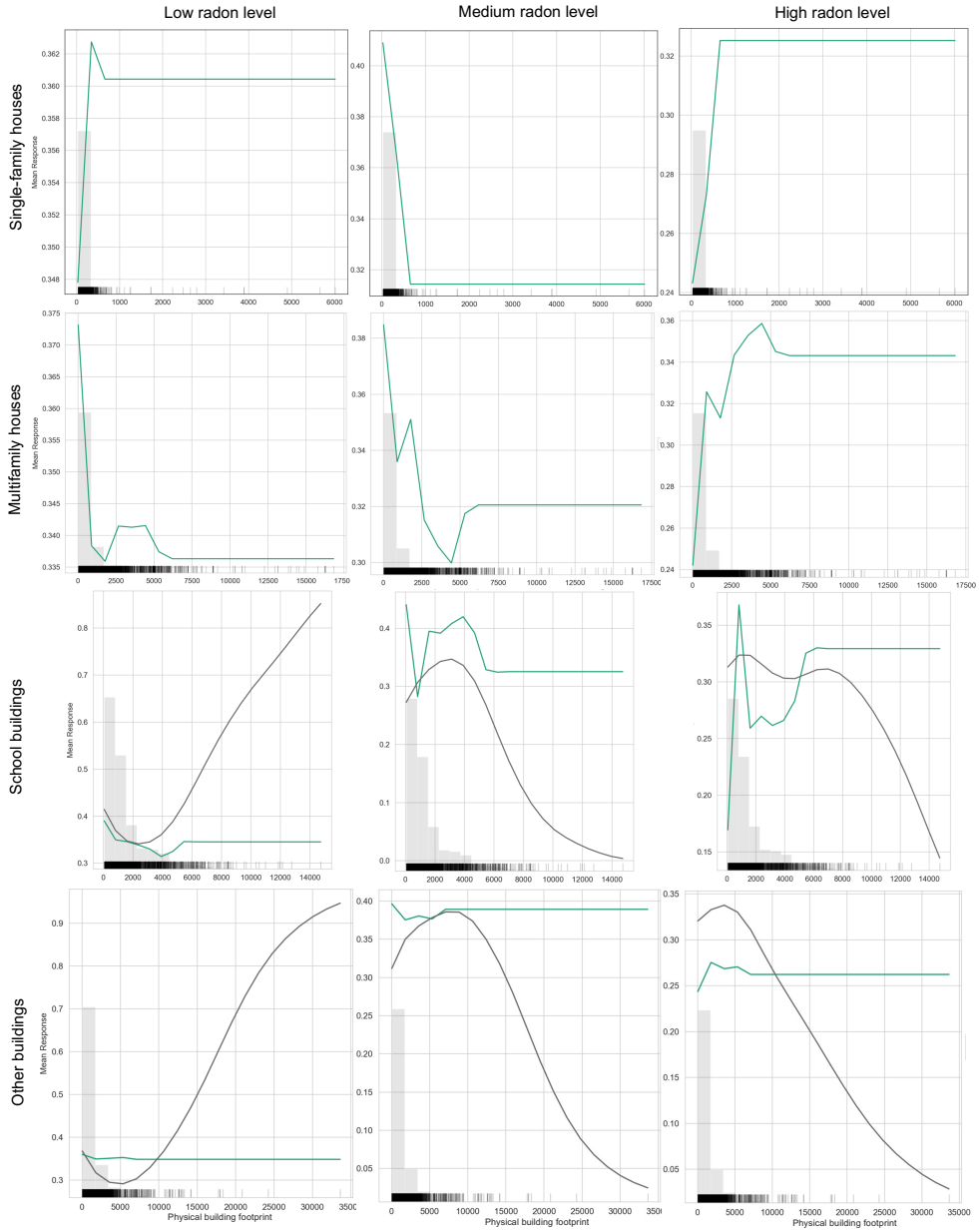
(ii) Asbestos-containing materials

Figure A1(i-ii). SHAP summary plots of lead models for hazardous materials prediction in residential and non-residential buildings. Geographically, the initial digit of postcodes in Swedish cities corresponds to specific regions: Stockholm is represented by 1, Malmö by 2, Gothenburg by 4, and Kiruna by 9. Regarding the building type codes, educational facilities are classified under codes 8 and 21. Single-family houses fall within the range of 30-32 and 35, while multifamily houses are categorized as 33. Industrial buildings are identified with codes 40-53, and office or commercial structures are marked as 99.

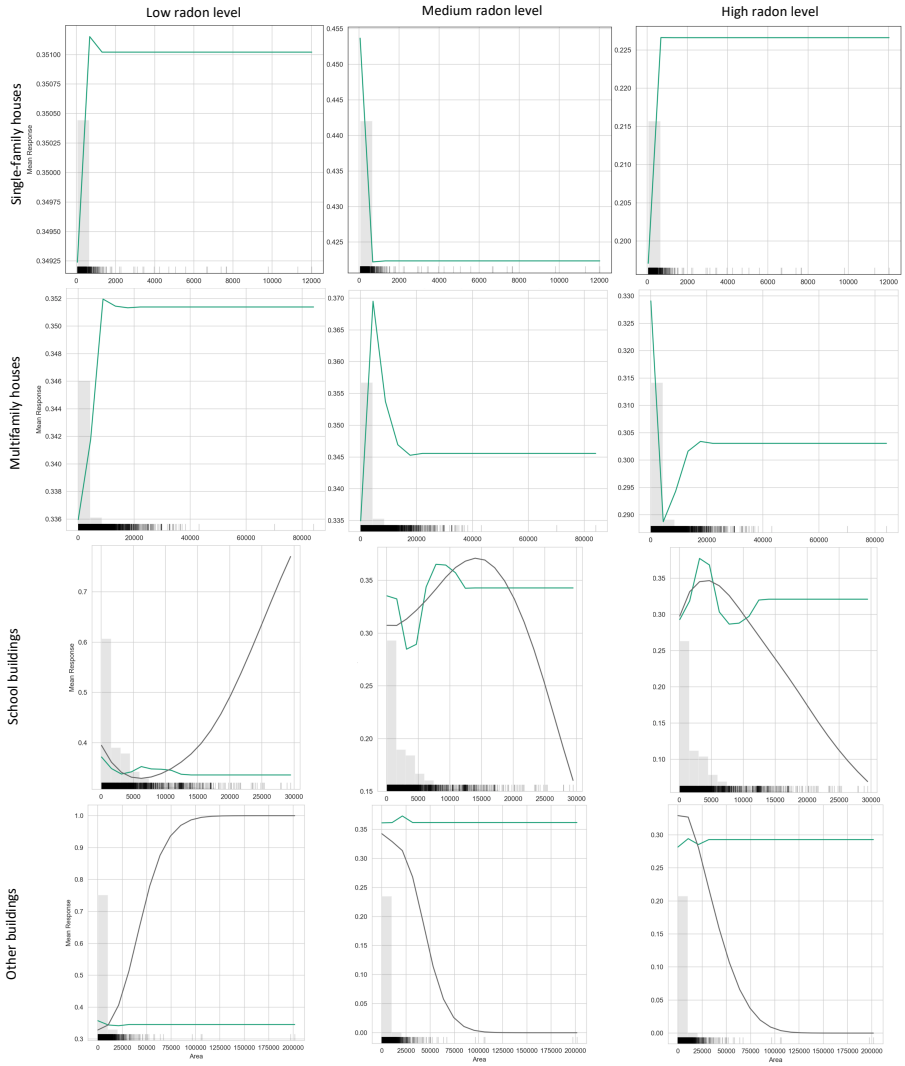
Appendix III Partial Dependent Plots for Indoor Radon Level Prediction



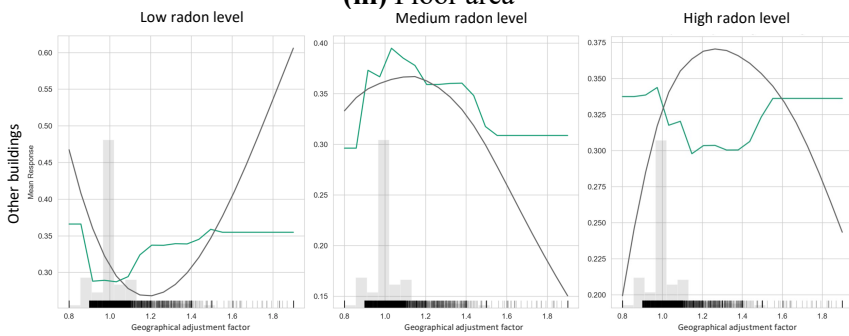
(i) Construction year



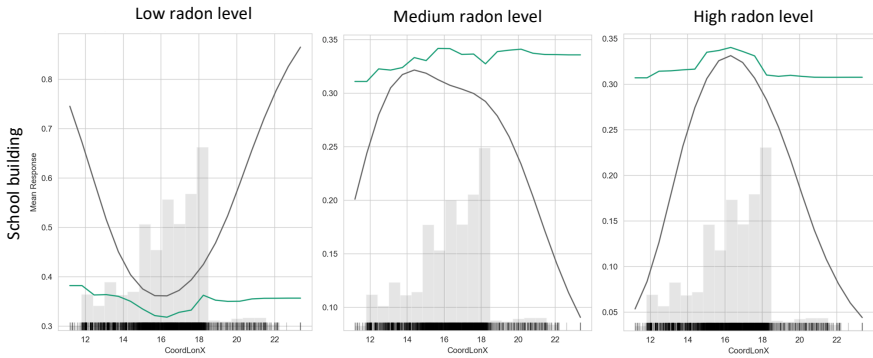
(ii) Building physical footprint



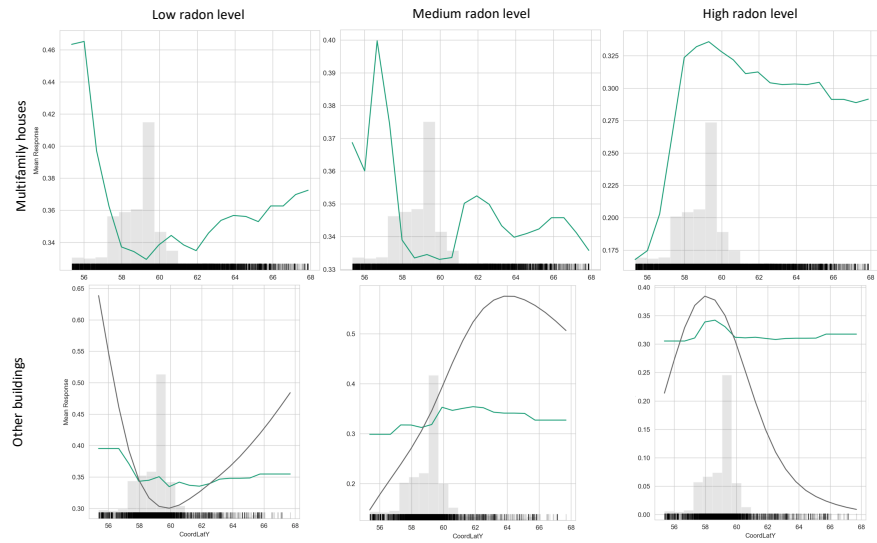
(iii) Floor area



(iv) Geographical adjustment factor



(v) Coordinate longitude



(vi) Coordinate latitude

Figure A2(i-vi). Partial dependent plots for indoor radon level prediction by building classes and features. The green lines representing the results from the XGBoost models, and black lines indicating those from the DNN models.

Tackling hazardous materials in circular construction

Alarming evidence from cutting-edge research has shed light on a concerning reality – six out of nine planetary boundaries have been breached with irreversible consequences looming on the horizon. While there's a resounding call to reduce our anthropogenic impact on Earth's systems, the construction industry seems to march in the opposite direction, continuously consuming more resources. To address this challenge of "building better from less," the construction industry is setting its sights on embracing circular practices. This transition is seen as an inevitable path to reduce the industry's climate impact. However, the road to integrating circular strategies – regenerating, narrowing, slowing, and closing resource loops – into current construction practices is slow and laden with challenges. One major obstacle lies in the misconceptions and misperceptions that pervade the construction industry regarding the costs, benefits, and feasibility of circular building practices. The thesis addresses the risk of unexpected encountering of hazardous substances in existing building stock by proposing predictive models as a decision support to assist stakeholders in estimating the presence of hazardous substances.

The doctoral thesis is composed by Pei-Yu Wu in collaboration with Lund University and RISE Research Institutes of Sweden. With a cross-disciplinary educational background from Architecture in Bachelor and Design and Construction Project Management in Master, Pei-Yu brings years of relevant professional experience to her research. Now, she channels her passion and expertise into the realms of circular construction and digital built environments, harnessing the power of applied AI to push the boundaries of her field.

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Engineering
Department of Building and Environmental Technology
Division of Building Physics
ISBN 978-91-88722-82-9
ISRN LUTVDG/TVBH--24/1030--SE (325)
ISSN 0349-4950