LUND UNIVERSITY

Estimating the probability distributions of radioactive concrete in the building stock using Bayesian networks

Wu, Pei-Yu; Johansson, Tim; Mangold, Mikael; Sandels, Claes; Mjörnell, Kristina

Published in:
Expert Systems with Applications

DOI:
10.1016/j.eswa.2023.119812

2023

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

Link to publication

Citation for published version (APA):
Wu, P.-Y., Johansson, T., Mangold, M., Sandels, C., & Mjörnell, K. (2023). Estimating the probability distributions of radioactive concrete in the building stock using Bayesian networks. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2023.119812

Total number of authors:
5

Creative Commons License:
CC BY

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

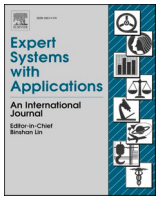
PO Box 117
221 00 Lund
+46 46-222 00 00

Expert Systems With Applications 222 (2023) 119812

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Estimating the probability distributions of radioactive concrete in the building stock using Bayesian networks

Pei-Yu Wu [a,b,*], Tim Johansson [a], Mikael Mangold [a], Claes Sandels [a], Kristina Mjörnell [a,b]

[a] RISE Research Institutes of Sweden, 412 58 Gothenburg, Sweden
[b] Department of Building and Environmental Technology, Faculty of Engineering, Lund University, 221 00 Lund, Sweden

ARTICLE INFO

Keywords:
Radioactive concrete
Building stock
Predictive inference
Bayesian network
Methodology
Risk-based inspection

ABSTRACT

The undesirable legacy of radioactive concrete (blue concrete) in post-war dwellings contributes to increased indoor radon levels and health threats to occupants. Despite continuous decontamination efforts, blue concrete still remains in the Swedish building stock due to low traceability as the consequence of lacking systematic documentation in technical descriptions and drawings and resource-demanding large-scaled radiation screening. The paper aims to explore the predictive inference potential of learning Bayesian networks for evaluating the presence probability of blue concrete. By integrating blue concrete records from indoor radon measurements, pre-demolition audit inventories, and building registers, it is possible to estimate buildings with high probabilities of containing blue concrete and encode the dependent relationships between variables. The findings show that blue concrete is estimated to be present in more than 30% of existing buildings, more than the current expert assumptions of 18–20%. The probability of detecting blue concrete depends on the distance to historical blue concrete manufacturing plants, building class, and construction year, but it is independent of floor area and basements. Multifamily houses and buildings built between 1960 and 1968 or nearby manufacturing plants are more likely to contain blue concrete. Despite heuristic, the data-driven approach offers an overview of the extent and the probability distribution of blue concrete-prone buildings in the regional building stock. The paper contributes to method development for pattern identification for hazardous building materials, i.e., blue concrete, and the trained models can be used for risk-based inspection planning before renovation and selective demolition.

1. Introduction

Existing buildings containing numerous hazardous materials cause health concerns for occupants and demolition workers (Kim & Yu, 2014). Blue concrete, a type of radioactive aerated concrete material, is one of the prominent examples associated with increased levels of indoor radon and heavy metals in buildings (Clavensjö & Åkerblom, 2020). The legacy of blue concrete in the Swedish building stock dates to the massive housing production between 1941 and 1975 across the country (Hall & Vidén, 2005). The alum shale, with high uranium and relatively low thorium contents, was used as fuel for lime firing and producing aerated concrete elements (Jelinek & Eliasson, 2015). As the uranium decays, the alum slate-based blue concrete and the ballast uranium-rich granite release 50–200 Bq/m^{2}h radon gas and emanate gamma radiation with radium. The radium content in blue concrete is 13–30 times more than ordinary concrete and releases 20–25 times more radon gas, making blue concrete a health hazard to be reckoned with in the indoor environment (Clavensjö & Åkerblom, 2020). Statistical results from radon measurement records show that an average radon level is 63,1% higher in Swedish buildings built with blue concrete than with ordinary concrete, depending on the amount of radium the lightweight concrete contains and the extent of usage in construction (Khan et al., 2021). The results from the early ELIB study (Sedin & Hjelte, 2004) also confirmed the increase of radon concentration by 10% in single-family houses and 20% in multifamily houses built with blue concrete. The presence of blue concrete in combination with ground source radon can further contribute to extreme indoor radon concentrations of more than 1000 Bq/m^3 if the ventilation system is poor (Clavensjö & Åkerblom, 2020).

Due to its lightweight characteristic and material availability, blue

* Corresponding author at: Sven Hultins Plats 5, 412 58 Gothenburg, Sweden.
E-mail addresses: pei-yu.wu@ri.se (P.-Y. Wu), tim.johansson@ri.se (T. Johansson), mikael.mangold@ri.se (M. Mangold), claes.sandels@ri.se (C. Sandels), kristina.mjornell@ri.se (K. Mjörnell).

https://doi.org/10.1016/j.eswa.2023.119812
Received 31 October 2022; Received in revised form 17 February 2023; Accepted 5 March 2023
Available online 9 March 2023
0957-4174/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

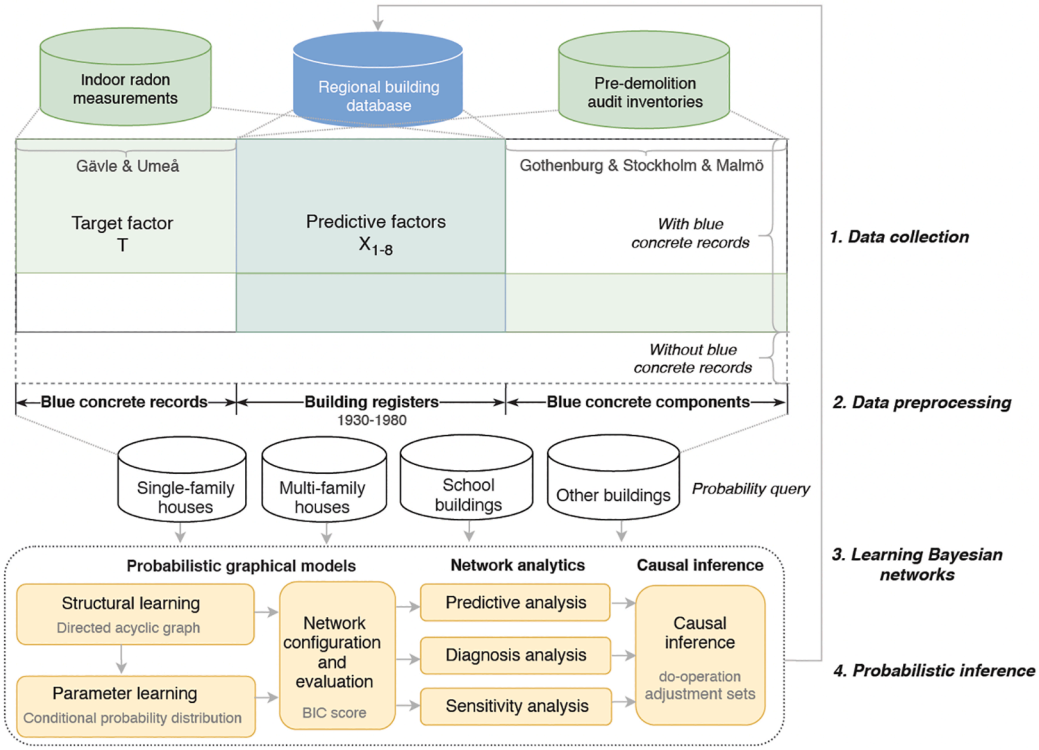


Fig. 1. Study outline for the predictive inference of blue concrete.

concrete was used extensively as reinforced and non-structural components in post-war dwellings. Blue concrete can be detected in walls and floor elements in single-family houses, external walls, and non-load-bearing partitions in multifamily houses (Clavensjö & Åkerblom, 2020). Even though the production and the use of blue concrete in construction had ceased for nearly half a century, it is estimated that blue concrete still exists in 6–7% of the existing building stock. Nowadays, blue concrete is investigated with a gamma radiation detection device if the results from the indoor radon measurements are above the reference level of 200 Bq/m^3. Buildings suspected of blue concrete are also obliged to conduct indoor radon measurements to ensure the radon concentrations are within the acceptable interval (Swedish Radiation Safety Authority, 2013). Otherwise, the blue concrete inspection will be carried out in the pre-demolition audits before major renovation or demolition due to the requirement of tremendous efforts for decontamination (Swedish National Board of Housing Building and Planning, 2010a, Swedish National Board of Housing Building and Planning, 2010b). In regard to this, risk-based inspection in predictive maintenance offers a cost-efficient alternative in the early identification and screening of blue concrete for large building stock (Bouabdallaoui, Lafhaj, Yim, Ducoulombier, & Bennadji, 2021).

The low availability of blue concrete registers is another barrier to tracing their presence in buildings. Currently, the information on blue concrete-containing components scatters among municipalities' databases and does not systematically connect with building registers. The geophysical flight measurement map of uranium also provides a hint of blue concrete, yet it cannot be used to ascertain blue concrete at the individual building level due to the mixed signals from ground radon (Jelinek & Eliasson, 2015). Nevertheless, the knowledge gap on the presence of blue concrete can be overcome by developing a data-driven approach for risk-based inspection. By coupling blue concrete records from past pre-demolition audit inventories and indoor radon measurements with the national building registers, it is possible to improve the traceability of specific hazardous materials on a regional or national scale and verify experienced-based expert assumptions (Wu, Sandels, Mjörnell, Mangold, & Johansson, 2022).

Learning Bayesian networks are one of the prediction methods with multifaceted benefits for building stock analyses. Developed from Bayes' Theorem, Bayesian networks are statistical learning tools widely used in building and environmental engineering disciplines to evaluate and ratiocinate uncertainty in risk assessments (Chen & Zhang, 2021). State-of-the-art literature ascertains Bayesian networks' applications in disaster risk analysis for building damage prediction under earthquakes (Chen & Zhang, 2021), fire hazard analysis in urban buildings (Liu, Lu, Xia, Li, & Zhang, 2017), uncertainty quantification in building inspection and diagnosis (Pereira et al., 2021), and probabilistic performance evaluation for building status (Bortolini & Forcada, 2020), etc. The modeling approach is reported to be effective for inferences despite incomplete information in building stock evaluation (Carbonari et al., 2019), which makes it a suitable instrument for pattern identification. Compared to other data-driven methods, i.e., machine learning or deep learning, Bayesian networks produce prediction outcomes with higher interpretability concerning descriptive, predictive, and prescriptive dimensions (Chen & Zhang, 2021). The probabilistic graphical models offer a descriptive overview of multivariate correlation with less demanding requirements for feature selection. The conditional probability distributions can be transformed into causal networks for diagnostic purposes by unfolding the rationality of the learned models and the causal relationships among input factors. Moreover, the networks trained on the sample population can be effectively transferred to perform probabilistic inference on the entire population. Considering these advantages and the previously mentioned data limitation, Bayesian networks are chosen in the study to predict the presence probability of blue concrete in the context of building stock.

As the first study investigating the applicability of Bayesian networks for in situ building material prediction, the paper aims to explore the following aspects: (i) characterizing the presence of blue concrete and containing components in various building classes and describing its correlations with measured indoor radon levels, (ii) constructing and transforming Bayesian networks to causal graphical models to untangle

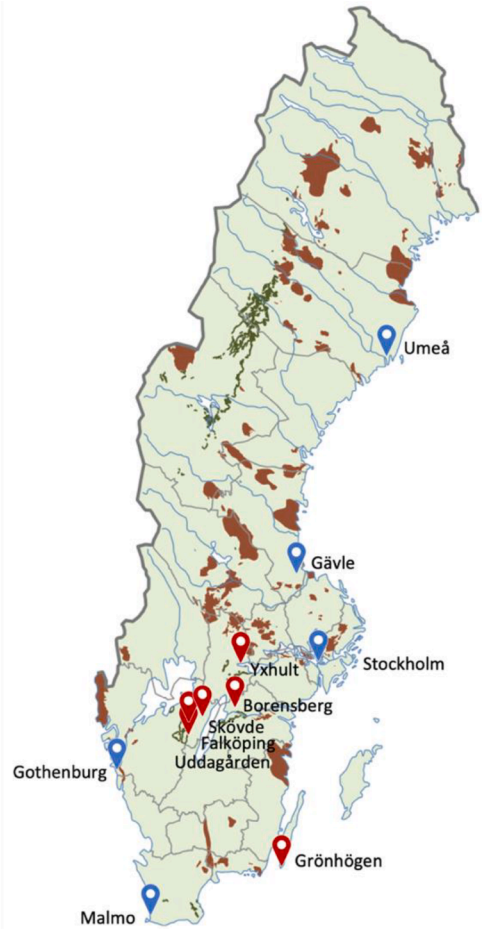


Fig. 2. Geographical map of areas with the presence of radioactive rock (brown) and regions with alum slate (dark green), adapted from the report by the Geological Survey of Sweden (Jelinek & Eliasson, 2015). The location of the five municipalities with blue concrete records (blue) and historical blue concrete manufacturing plants (red) are annotated.

relationships between factors, and (iii) estimating the extent of building stock with a high probability of containing blue concrete by applying predictive inference to regional building registers. Developing the predictive method for hazardous material risk evaluation, in terms of probability estimation and causality inference, can facilitate environmental inspection and remediation planning (Kim, Hamann, Sotiralis, Ventikos, & Straub, 2018). The prerequisite of developing the integrated approach is built upon a rather homogeneous building stock where buildings were constructed with similar dimensions, typologies, morphologies, and construction features (Lucchi, Exner, & D'Alonzo, 2018). The residential stocks in Sweden built between 1945 and 1975 conform to this requirement, and the systematic-built housings can be categorized into building types based on building characteristics (Björk, Kallstenius, & Reppen, 2013). The study outcomes can be used as decision support when pinpointing buildings with high intervention priorities for blue concrete decontamination.

2. Material and methodology

2.1. Study design

The process of constructing Bayesian network models for blue concrete prediction and inference to the Swedish building stock comprises four parts, depicted in Fig. 1 below.

The first part of the study concerns data collection, where building-generic data, i.e., building registers, and building-specific data, i.e., indoor radon measurements and pre-demolition audit inventories, were compiled. Afterward, the collected data underwent several preprocessing steps, including data integration, cleaning, discretization, and node selection before modeling. The third part involves constructing two types of networks – structural leaning and parameter learning – and verifying model performance with various scoring metrics. Then network analytics and causal inference were conducted to evaluate the developed Bayesian network models and improve the interpretation of the results. Lastly, the models were applied to the regional building dataset from the five municipalities to query the probabilistic distributions of the remaining blue concrete.

2.2. Material

The input data in the study constitute building-specific material data and national building registers in Sweden. Integrating the inspection records of blue concrete (target variable) and the national building registers (predictive variables or label instances) formed the foundation of the blue concrete dataset, described in Appendix A.

2.2.1. Data sources and compilation

The information on blue concrete was assembled from two data sources – the municipality indoor radon measurements and the pre-demolition audit inventories – for buildings built between 1930 and 1980. The municipality's indoor radon measurements contain yearly average indoor radon levels from trace film measurements and the occurrence of suspected blue concrete in buildings by radiation scanning vehicles, while the pre-demolition audit inventories offer more detailed information on blue concrete components. Compiling these accessible datasets enables us to analyze the presence of blue concrete at building and component levels. In the study, open indoor radon datasets from Gävle and Umeå municipalities in Sweden were retrieved from the Swedish data portal maintained by the Agency of Digital Government (DIGG, 2022). Houses built with blue concrete were mapped in early state initiatives of scanning radiation from vehicles in the Swedish municipalities (Statens offentliga utredningar från Näringsdepartementet, 2001; Statens offentliga utredningar, 1983). These datasets include 2,831 blue concrete inspection records for mainly residential buildings. Simultaneously, detection records of blue concrete components were assembled from pre-demolition audit inventories from renovation and demolition projects in major Swedish cities, Gothenburg, Stockholm, and Malmö, whose metropolitan regions comprise 48% of the heated floor area and around 35% of the total number of buildings in the Swedish multifamily housing stock (Björk et al., 2013). 325 observations from various building classes containing blue concrete in different components were retrieved from pre-demolition inventories. The geographical locations of the municipalities, historical blue concrete manufacturing plants, radioactive rock, and regions with alum slates are illustrated in Fig. 2. It is worth noticing that the five municipalities in the study are not directly situated on radioactive rock or regions with alum slates.

The national building dataset was compiled from the Swedish Energy Performance Certificates (EPCs), the municipality cadastral register, and the building taxation register to present the entire building stock. These registered data comprise comprehensive information on building usage, i.e., building category and types, and building parameters, i.e., construction year, floor area, number of basements and floors, etc. Furthermore, the collected blue concrete data from inspection records and radon measurements indicated blue concrete was merged and matched with the building registers using the national real estate index and address as matching keys in FME (Feature Manipulation Engine) from Safe Software. Considering the substantial construction period of blue concrete, a subset of the buildings built between 1930 and 1980 was selected for predictive inference. In total, 2,424 observations remain for subsequent data analysis and learning Bayesian Networks modeling.

Table 1
Value distribution and mean values of building parameters from the data subset of buildings with blue concrete records constructed between 1930 and 1980 based on building classes (N = 2,424).

Building parameters	Building classes			
	Residential building		Non-residential building	
	Single-family house	Multifamily house	School building	Other building
	n = 1,841	n = 312	n = 101	n = 170
Construction year	1930–1980 (1963)	1930–1980 (1960)	1933–1980 (1965)	1930–1980 (1961)
Floor area (m^2)	40–640 (190)	232–38,230 (2188)	70–48,462 (2935)	100–111,097 (8057)
Number of floors	1–3(2)	1–14(3)	1–7(2)	1–26(4)
Number of basements	0–1(1)	0–2+(1)	0–2+(1)	0–2+(2)
Number of stairwells	0–1(0)	0–19(2)	0–9(1)	0–8(2)
Number of apartments	0–8(1)	0–341(23)	0–2(0)	0–376(10)
Blue concrete detected	257 (14%)	165 (53%)	25 (25%)	83 (49%)

2.2.2. Statistical description of the observed buildings

Clustering buildings with similar characteristics is critical to partition building stock into comparative and representative typologies (Lucchi et al., 2018). Data stratification is crucial for applying inference from analytical results to other instances with similar data profiles accurately. According to primary usage and building dimensions, the building stock can be categorized into residential and non-residential buildings. Then based on the municipality building usage code, residential buildings are categorized into single-family and multifamily houses. Non-residential buildings are more complex to be generalized to a specific genre. School buildings in Sweden are built and operated by municipalities and thus have rather similar technical details on a regional basis, while the rest of the building classes, such as commercial buildings, office buildings, and industrial buildings, are categorized as other buildings. To address the sample representation, floor areas were retrieved from the latest building taxation data to compare the regional building stock from the five municipalities and the national building stock. The statistic shows that the five municipalities' living areas represent 8% of the entire single-family houses, 31% of multifamily houses, and 16% of office buildings. School buildings are exempted from taxation and are left out of the building taxation registers.

Furthermore, by stratifying data with building classes, an overview of numerical building parameters for each building class was obtained, which could be helpful for understanding the data structure and identifying data noise or outliers. As shown in Table 1, blue concrete and the typical components containing blue concrete are detected in 14% of single-family houses, 53% of multifamily houses, 25% of school buildings, and 49% of other buildings in the building stock 1930–1980. The average construction years of these buildings are from the 60s, corresponding to the historical timeline of the construction peak in Sweden. Besides, the average floor areas are distinctive between building classes, which could be indicative of variable discretization. Overall, the average values of building parameters from residential and non-residential buildings agree with the existing knowledge of the building stock (Björk et al., 2013) and thus can be assumed to be representative of the uninspected building stocks built in the same period.

2.3. Methodology

The method section starts with a description of the theoretical background of risk-based inspection and Bayesian networks. Then a sequential workflow – data preprocessing, Bayesian network modeling, and predictive inference – adopted in the study were illustrated.

2.3.1. Theoretical background

Decision support tools, such as statistical learning models, facilitate the identification of renovation and maintenance strategies to enhance the conservation state of buildings (Bortolini & Forcada, 2020). These data-driven models provide the scientific basis to evaluate potential technical solutions according to the input data from historical records. Nevertheless, modeling hazardous materials on an urban scale involves uncertainties concerning empirical data quality, i.e., completeness, consistency, and accuracy (Wu, 2022; Wu, Mjörnell, Mangold, Sandels, & Johansson, 2021). The uncertainty relates to the building inspections and diagnosis of the building status and elements inherent to the subjectivity of surveyors (Pereira et al., 2021). Therefore, former research has tried to address the uncertainties by engaging risk-based inspections with Bayesian networks in civil engineering disciplines (Carbonari et al., 2019; Kim et al., 2018; Liu et al., 2017). The developed probability models for defects or hazard detection are used to prioritize inspections to achieve optimal planning of time and resources.

2.3.1.1. Risk-based inspections. Risk-based inspections refer to the processes of developing a scheme of inspection to evaluate the probability and consequences of defects (Kim et al., 2018). It involves qualitative and quantitative assessment of the likelihood of failure and the consequence of failure (Kim et al., 2018). Compared to traditional rule-based or condition-based inspections, risk-based inspections are more efficient in specifying the inspection scope for particular parts with higher risk (Kim et al., 2018). Risk-based inspections enable us to bring forward a risk-based building retrofit planning framework that is economical and feasible (Carbonari et al., 2019). A specific example is the adoption of the SOBANE strategy (screening, observation, analysis, expertise) in the building sector for risk management (Lucchi, 2016): (i) screening the performance hotspots; (ii) observing the detecting causative factors associated with presenting and potential risk; (iii) analyzing and quantifying the environmental risks with field investigations and measurements; (iv) expertizing guidelines for solutions prioritization and implementation for conservation or renovation. Identifying uncertain factors affecting building conditions and their relationships can support decision-making from holistic and realistic perspectives (Bortolini & Forcada, 2020). Bayesian networks are one of the quantitative techniques for forecasting the probability of failure in risk-based inspections when information is limited, accessible, or incomplete (Bortolini & Forcada, 2020).

2.3.1.2. Bayesian networks. Bayesian network is a probabilistic graphical model that factorizes the joint distribution of variables and represents their interdependent relationships (Chen & Zhang, 2021; Liu et al., 2017). The probabilities of variables are non-static with the addition of new observations; thus, the models are useful for predictive analytics and inference to identify the effects and strength of relationships between variables (QuantumBlack, 2020). In Bayesian networks, the information is structured in the format of Directed Acyclic Graphs (DAGs) to describe dependencies (edges) between connecting input factors (nodes) (Chen & Zhang, 2021). The joint probability distribution of all factors of the Bayesian networks can be denoted in Eq. (1), where a set of factors is represented as V = {Xi | i = 1, 2, …, N} and the set of parent factors of the ith node represented as Pr(Xi) (Chen & Zhang, 2021). *N* signifies the total number of factors over a set of terms of the conditional distribution function of *Xi*.

$$P(X_1, X_2, \ldots, X_n) = \Pi_{i=1}^{N} P(X_i | Pr(X_1)) \quad (1)$$

Probabilistic or causal inferences are performed by updating the posterior probabilities of target nodes with evidence from input nodes under the given circumstances, as shown in Eq. (2). The posterior probability of target variable $P(T|E)$ with given evidence can be calculated by aggregating the joint probability of T and E and the prior probability of E.

Table 2
Data discretization for the buildings built between 1930 and 1980 with blue concrete inspection records.

Node	Factor	Data representation [nominal / ordinal]	N [NA %]
X_1	Construction year	year [1930–1955, 1955–1960, 1960–1968, 1968–1974, 1974–1980]	2424 (0%)
X_2	Floor area	m^2 [0–150, 150–220, 220–360, 360–1500, over 1500]	1150 (53%)
X_3	Building class	Nominal [Single-family house, Multifamily house, School building, Other building]	2424 (0%)
X_4	Average distance (to manufacturing plants)*	km [0–300, 300–600, over 600]	2424 (0%)
X_5	Basements	Count [no basement, at least one basement]	756 (69%)
X_6	Number of floors	Count [1, 2, 3, 4–6, over 6]	724 (69%)
X_7	Number of stairwells	Count [0, 1, 2, 3, over 4]	607 (75%)
X_8	Number of apartments	Count [0, 1–2, 2–10, 10–30, over 30]	587 (76%)
T	Blue concrete	Nominal [positive, negative, NA]	2424 (0%)

* Historical blue concrete manufacturing factories in Sweden and their operating period are Borensberg (1936–1968), Yxhult N:a (1929–1959), Yxhult K (1966-), Yxhult S:a (1947–1975), Falköping (1930–1974), Uddagården (1955–1974), Grönhögen at Öland (1943–1972), and Skövde/Durox (1925–1968) according to the Swedish Radiation Protection Authority (Clavensjö & Åkerblom, 2020). The average distances were computed between the city center of the municipality where observed building and manufacturing plants were situated and then categorized into three distance groups.

$$P(T|E) = \frac{P(T,E)}{P(E)} \tag{2}$$

Bayesian network learning is two-fold, depending on the expert knowledge of relationships between target and predictive variables. In the case of unknown causal relationships, structural learning can be applied to construct reasonable DAGs from the given data based on conditional independence tests of factors (Chen & Zhang, 2021). By screening the optimal DAGs, a data-driven graph and directions of edges close to reality can be generated. On the other hand, *parameter learning* is used to identify the conditional probability distribution of nodes in predefined DAGs. The learning process leverages the maximum likelihood estimation to fit the data. Yet, the risk of overfitting exists, and the trade-off between overfitting and fitting by imposing prior distribution should be considered (Gao et al., 2019). The direction of and joint probability distribution of blue concrete remains unclear. Thus, structural learning will be applied as the first step to creating DAGs, followed by computing a conditional probability table (CPT) for each node using parameter learning. Then predictive inference of blue concrete in unknown building stock can be performed based on the trained and evaluated Bayesian network models.

2.3.2. Data preprocessing

Since the blue concrete inspection records were collected from multiple sources, data cleaning, including terminology harmonization and missing value imputation, was performed to create a coherent, machine-readable dataset. Concerning the different levels of details from building-specific data, the subset of the pre-demolition audit inventories was retrieved to analyze the frequent presence of blue concrete-containing components in the building stock, including the wall, floor, façade, ventilation shaft, others, and unspecific. Others are, for example, roof or fire cell prohibition walls in the attic. A few buildings were detected with multiple blue concrete-containing components and labeled as multiclassification observations. Afterward, missing values of building parameters from registers, such as floor area, number of floors, and basements, were constructed by examining their Google Street View images, plan drawings and inventory reports. As the geographical location of manufacturing plants is shown to be relevant to the spatial distribution of hazardous building materials in a previous study (Wilk, Krówczyńska, & Zagajewski, 2019), the average distance to blue concrete manufacturing factories was computed between the city center of the municipality where the observed building is situated and the historical locations of the factories as an additional feature.

After that, preliminary node selection is performed to remove redundant features in a dataset to prevent bias and improve classification accuracy. Node selection criteria are based on expert knowledge (Rönnqvist, 2021) and literature (Boverket, 2013; Clavensjö & Åkerblom, 2020) regarding the potential causal factors of the presence of blue concrete, including geographical attributes and building characteristics. Then binning technique was applied to the factors to generate three to five discrete intervals containing similar numbers of observations. The data discretization of splitting continuous variables into data subgroups is a prerequisite, as the Bayesian network algorithms are only compatible with discrete variables. An overview of available data amount and representation for selected factors are presented in Table 2.

2.3.3. Bayesian network modeling

The Bayesian network modeling was facilitated by Python Bayesian libraries that contain extensive pipelines for Bayesian network learning and inference (Ankan & Panda, 2015). Structural learning was performed in the first part of model development to explore unknown relationships or dependencies between factors for DAG (pattern) creation. Various structure learning algorithms were investigated in search of the best DAG fitting to the given data, including Exhaustive Search, Hill-Climb Search, Tree Search (Chow-Liu), PC (Constraint-based estimator), and Max-Min Hill-Climb. Taking score-based search strategies, for instance, the Exhaustive Search is suitable for small networks with less than five nodes but challenging to identify the ideal structure due to lacking local optimization. In contrast, the other heuristic search approach Hill-Climb Search, can handle more nodes by executing a greedy local search iteratively until a local maximum is found (Taskesen, 2022). On the other hand, the Tree Search algorithm can operate on massive datasets involving complicated uncertainties among various interdependent feature sets (Taskesen, 2022). Another way is to use constraint-based structural learning, such as the PC estimator, to identify independencies in the dataset using hypotheses (Ankan & Panda, 2015). The hybrid method combining score-based and constraint-based structure learning algorithm Min-Max Hill-Climb estimates the graph skeleton with PC and then orients the edges using hill-climb search.

Based on the variable dependencies of the networks, three scoring functions were employed to evaluate the probabilistic models' performance. *K2* metric assumes a uniform prior distribution on the values of a node for each possible instantiation of its parent nodes and one to the count of every state, which makes it tend to choose simplified networks (Borgelt & Kruse, 2001). On the other hand, *BDethe u* metric *(Bayesian Dirichlet equivalent uniform prior*) uses observed N uniform samples of each variable as pseudo-counts and is sensitive to parameter settings. In comparison, *Bayesian Information Criteria (BIC)* score is a relatively robust scoring metric and is reported to outperform the BDeu metric in empirical studies (Liu, Malone, & Yuan, 2012). As defined in Eq. (3), the BIC metric regulates model complexity by introducing a penalty under the maximum likelihood estimation (Ankan & Panda, 2015). This metric describes how well a model captured the underlying structure of the data and was used for model selection, of which a lower BIC value implies lower penalty terms and hence is preferable.

$$\mathrm{BIC} = \log(n)k - 2\log(\widehat{L}) \tag{3}$$

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

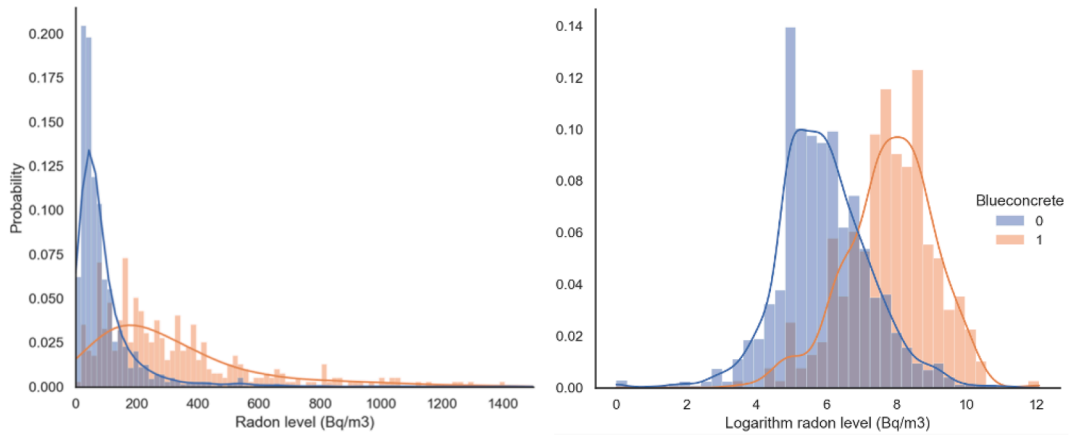


Fig. 3. Normalized density distribution of annual average radon levels (left) and logarithm radon levels (right) grouped by buildings with (N = 398) or without (N = 1,808) blue concrete detection for probability estimation.

Subsequently, parameter learning was carried out to estimate conditional probabilities distributions (CPDs) of individual variables. The type of parameter learning adopted in the study is *Bayesian Parameter Estimation*, which leverages existing prior CPDs (or pseudo-state counts) to the observed data to update the posterior. This method can overcome the risk of overfitting the data if the observed samples are not representative of the actual distribution or complete that occurs with *Maximum Likelihood Estimation* (Taskesen, 2022). In Bayesian Parameter Estimation, priors are usually set to be a constant value in every iteration to have equiprobable states, and then computed CPDs were related to the selected DAGs to construct Bayesian networks.

Furthermore, three types of network analytics were performed to gain a more holistic understanding of models' behavior under different circumstances. *Prediction analysis* was used to investigate the predictive performance of the Bayesian networks assuming various extents of evidence (Chen & Zhang, 2021). Given different data availability of variables in the building databases, three levels of evidence are considered: (i) Scenario I: X_1-X_3 (Construction year, Floor Area, Building class); (ii) Scenario II: X_1-X_6 (Construction year, Floor area, Building class, Distance to manufacture plants, Basements, Number of floors); (iii) Scenario III: full evidence X_1-X_8 (Construction year, Floor area, Building class, Distance to manufacture plants, Basements, Number of floors, Number of stairwells, Number of apartments). Scenario I is a baseline model containing variables available in all sorts of registers for any buildings in Sweden, including municipality cadastral registers, building taxation registers, and EPCs. Scenario II adds extra information on the geographical distances to blue concrete manufacturing plants and the number of floors and basements that are often available in pre-demolition audit inventories or plan drawings from building permit documents. Extensive information in scenario III can be retrieved from EPC, encompassing around 92% of multifamily houses (including commercial use), 25% of single-family houses, and most school buildings (Johansson, Olofsson, & Mangold, 2017).

The *diagnosis analysis* features backward reasoning to identify the most influential factor associated with the potential presence of blue concrete (Chen & Zhang, 2021). Derived from the Bayesian theorem, diagnosis analysis compares the changes in posterior probabilities for different factors by updating the state of the blue concrete detection in a stepwise manner. Lastly, sensitivity analysis was performed to uncover the factors contributing to the substantial variation of model outputs through changing intervals of discrete variables (Chen & Zhang, 2021). By measuring the sensitivity of the input nodes in the detection of blue concrete, susceptible factors can be highlighted for risk abatement beforehand. The results from these network analytics help to build a holistic understanding of the model uncertainty given various variables, identification of critical factors in the Bayesian networks, and sanity check for validity of test assumptions.

Proceeding with *causal inference*, the Bayesian networks were transformed into causal networks to identify conditional independencies using d-separation algorithms (Pearl & Dechter, 2013). The invariance of the structure of the models and the relationships between nodes was tested when intervention (*do-operation*) occurred (Barr, 2018). After that, the *backdoor adjustment formula* shown in Eq. (4) was applied to estimate the causal influence of X on Y given certain circumstances (W).

$$\mathrm{P}(Y|\mathrm{do}(X)) = \sum_{w} P(Y|X,W)\ \mathrm{P}(W) \tag{4}$$

2.3.4. Probabilistic inference

The probabilistic inference leverages the encoded probability distribution of the Bayesian network models to predict the presence patterns of blue concrete in regional buildings. By fitting the trained models to the building registers of the five municipalities, i.e., Stockholm, Gothenburg, Malmö, Gävle, and Umeå, the models perform queries with hard evidence to estimate the buildings potentially containing blue concrete in the regional building stock built between 1930 and 1980. Registers of the buildings constructed during this period were retrieved and processed as input data following the same procedure described in Sections 2.3.2 and 2.3.3. The same variable binning intervals were applied to ensure model transferability to the regional building dataset. Estimations were made by exploiting the generic Bayesian networks to gain an overview of the probabilistic distribution of the residual blue concrete in existing buildings. Also, data representativeness was assured through training the Bayesian networks on the observations from the population of the exact geographical locations. Afterward, the results of the probabilistic inference from different Bayesian models were compared to evaluate the generalizability and scalability of the approach.

3. Results

The presence of blue concrete and its components in buildings were investigated through statistics and data analysis. Afterward, Bayesian networks were developed and evaluated with probabilistic graphical models, network analytics, and causal inference. The last part concerns applying the Bayesian networks models to the blue concrete buildings with unknown blue concrete status in the regional building stock using probabilistic inference.

3.1. Characterization of blue concrete in buildings

To determine the impact magnitude of blue concrete on the indoor radon levels, the radon level distribution for buildings detected with and

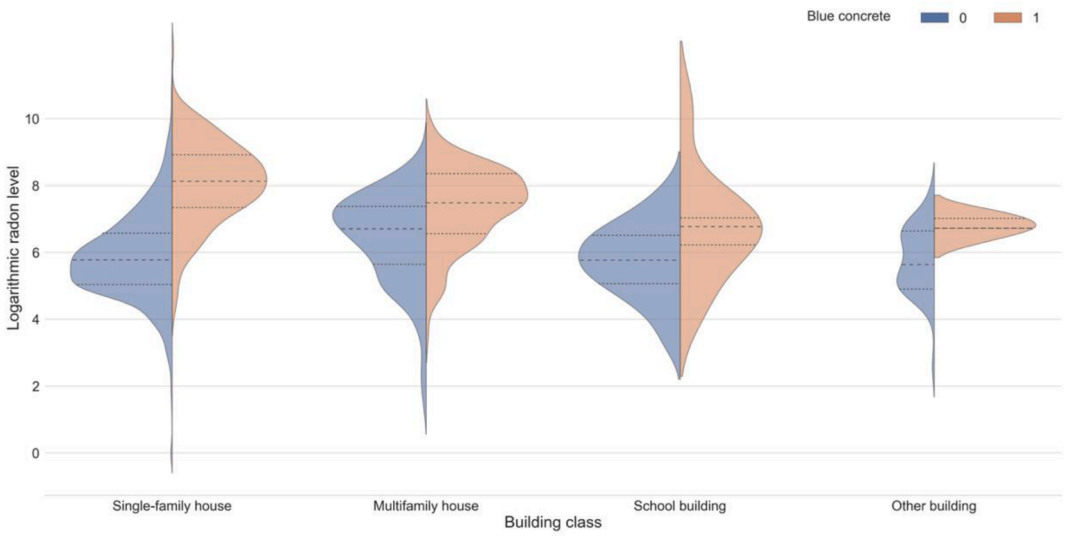


Fig. 4. Quartile distribution of the logarithmic radon level between buildings with and without blue concrete detection among building classes.

Table 3
Characterization of blue concrete and its components concerning annual average indoor radon levels in different building classes. The data count, arithmetic mean, and confidence interval (CI) of the radon levels (Bq/m^3) were computed for the blue concrete detection and non-detection buildings (N = 2,424, of which $N_{Blue\ concrete}$ = 2,206 and $NA_{Blue\ concrete}$ = 218).

Blue concrete component	Single-family house		Multifamily house		School building		Other building		Total	
	N	Bq/m^3	N	Bq/m^3	N	Bq/m^3	N	Bq/m^3	N	Bq/m^3
Walls	–	–	19	223 ± 58	3	98 ± 17	9	266 ± 264	31	223 ± 83
Fasade	–	–	3	260 ± 0	–	–	4	100 ± 20	7	169 ± 65
Floor/foundation	–	–	–	–	2	106 ± 0	2	305 ± 149	4	206 ± 192
Others	–	–	–	–	–	–	4	90 ± 39	4	90 ± 39
Unspecific	252	377 ± 46	106	215 ± 47	2	140 ± 20	3	277 ± 264	363	328 ± 34
Detection	252	377 ± 46	128	218 ± 26	5	115 ± 23	13	289 ± 191	398	319 ± 31
Non-detection	1,579	91 ± 7	135	119 ± 15	53	70 ± 12	41	161 ± 60	1,808	94 ± 6
Total	1,831	130 ± 9	263	167 ± 16	58	73 ± 11	54	192 ± 65	2,206	135 ± 8

*The highest acceptable annual average indoor radon level is 200 Bq/m^3 in Sweden.
** Confidence interval = sample mean ± margin of error (standard error).

without blue concrete was explored. The normalized density distribution in Fig. 3 estimates the probability of annual average and logarithm radon levels of the blue concrete dataset. The radon levels appear to be non-symmetric distributed with long tails in both subgroups, where approximately 180 Bq/m^3 marks the significant probability change. Buildings built without blue concrete are highly likely to be measured with a radon level below 100 Bq/m^3, whereas 200 Bq/m^3 or above is expected in buildings containing blue concrete. Further transforming the radon level to a logarithmic scale, the distributional discrepancy between the blue concrete subgroups is evidential and normally distributed. Accordingly, it is affirmative that blue concrete is closely associated with a higher radon level in buildings. The quantile distribution of the violin plots in Fig. 4 displays a more detailed spread of the logarithmic values by building classes. Despite various shapes, the blue concrete subgroups were found to have higher median values across building classes.

Table 3 characterizes the presence of blue concrete and its components in relation to radon. Blue concrete is present in several building parts, and the detection frequency of the containing components varies between building classes. Around 18% of buildings in the blue concrete dataset contain blue concrete, and they were found most frequently in interior walls, building facades, floor or foundation construction, and others, i.e., ventilation shafts. Blue concrete inspections conducted in pre-demolition audit inventory are less common for single-family houses, and thus the data is insufficient. Nevertheless, for complex or large-scale buildings, it was found frequently in walls and facades in multifamily houses and walls and floors or foundations in school buildings. Blue concrete was detected in several parts of other buildings, such as wall construction. Concerning the radon levels, the variance between average radon levels in residential dwellings was four times higher in single-family houses and almost doubled in multifamily houses. In school buildings, the average radon levels in blue concrete and non-blue concrete detected buildings were within acceptable levels, and the difference was minor. On the contrary, the radon level in other buildings is generally higher than the rest of the subgroups, and the variance level shows a similar pattern to multifamily houses.

Furthermore, the relationships between blue concrete, building typology, and indoor radon were investigated to identify predictive features. To untangle the interaction between variables, Pearson correlation matrixes in the heatmap were plotted in Fig. 5. The computed point-biserial correlation indicates coefficients and significance level between blue concrete and predictive variables for each building class. The average distance between observed buildings and historical blue concrete manufacturing plants has the foremost negative correlation to the presence of blue concrete in residential buildings, and the results are statistically significant. Construction year, floor area, basements, and the number of apartments and stairwells are also indicators for blue concrete in multifamily houses. Surprisingly, basements have opposite effects, reported a positive correlation to blue concrete in single-family houses and a negative in multifamily houses. Single-family houses built with basements show a slightly higher radon level (≈ 9% increase) than those without, whereas reverse situations

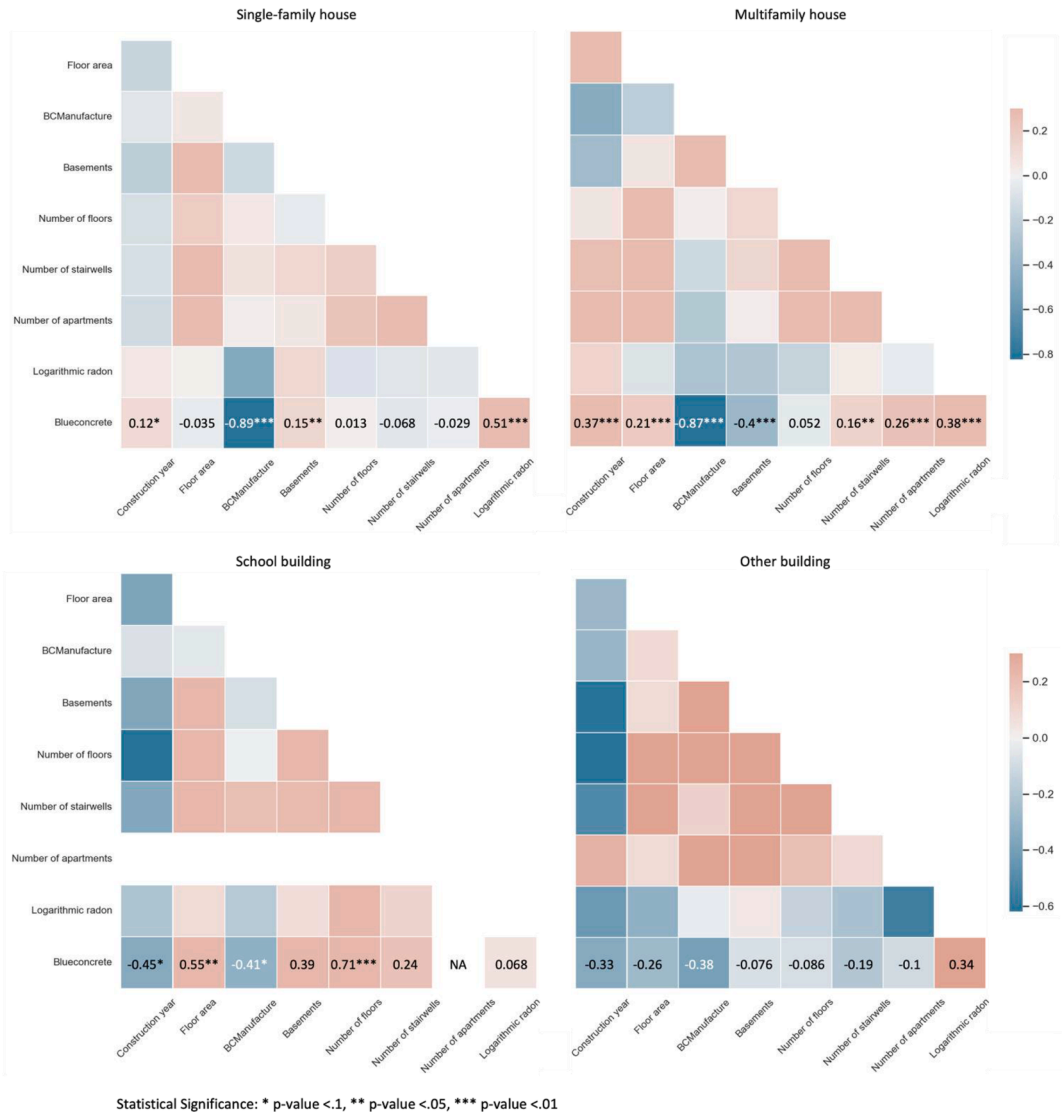


Fig. 5. Pearson correlation matrix based on point-biserial correlation indicating coefficients and significance level between blue concrete and predictive variables for each building class.

were observed in multifamily houses (≈ 30% decrease) and school buildings (≈ 6% decrease). The variation in radon levels may be due to the connectivity between basements and ground floors, where basements can become an open path for ground radon leakage in single-family houses. Yet, the basements in multifamily houses are usually not directly linked with other spaces. As for school buildings, the number of floors and floor area correlates positively to blue concrete. However, no variables are significantly associated with blue concrete in other buildings.

Appendix B compiles the hierarchical clustering of radon levels by potential variables. Data sufficiency was considered when computing the mean values and confidence intervals of radon levels to evaluate the validity of the results. The radon concentration at level 1 is regarded as a baseline when investigating the combined impacts of building parameters in each building class, i.e., basements and ventilation types. The results show that exhaust and balanced ventilation lower radon concentrations by around 17% and 36% compared to natural ventilation in multifamily houses. Similar trends were also observed in single-family houses, schools, and other buildings, but more data are required to validate the tendency. Essentially, buildings containing blue concrete were measured with significantly higher radon concentrations that can hardly be compensated by implementing ventilation measures. Buildings prone to high radon risk were outlined: single-family houses built with basements and blue concrete, multifamily houses built without basements but with blue concrete, and other buildings. The confidence intervals for single-family houses detected with blue concrete and other buildings are large, suggesting a higher uncertainty in representing the population mean.

3.2. Learning Bayesian networks for blue concrete

The Bayesian networks for blue concrete were constructed using structural learning and parameter learning, then evaluated with several scoring functions to find the optimal hyperparameter combination. Network analytics were performed to test the robustness of models and identify crucial variables in the prediction. Finally, causal inference was applied to the networks to configure relationships between variables and improve interpretability.

3.2.1. Probabilistic graphical models

In search for the optimal combination and number of nodes, extensive Bayesian models were constructed and trained on four structural learning algorithms, including *Hill-climb Search (hc), Tree Search (ts), Constraint-based estimator (PC), and Max-Min Hill-climb Search (mmhc)*.

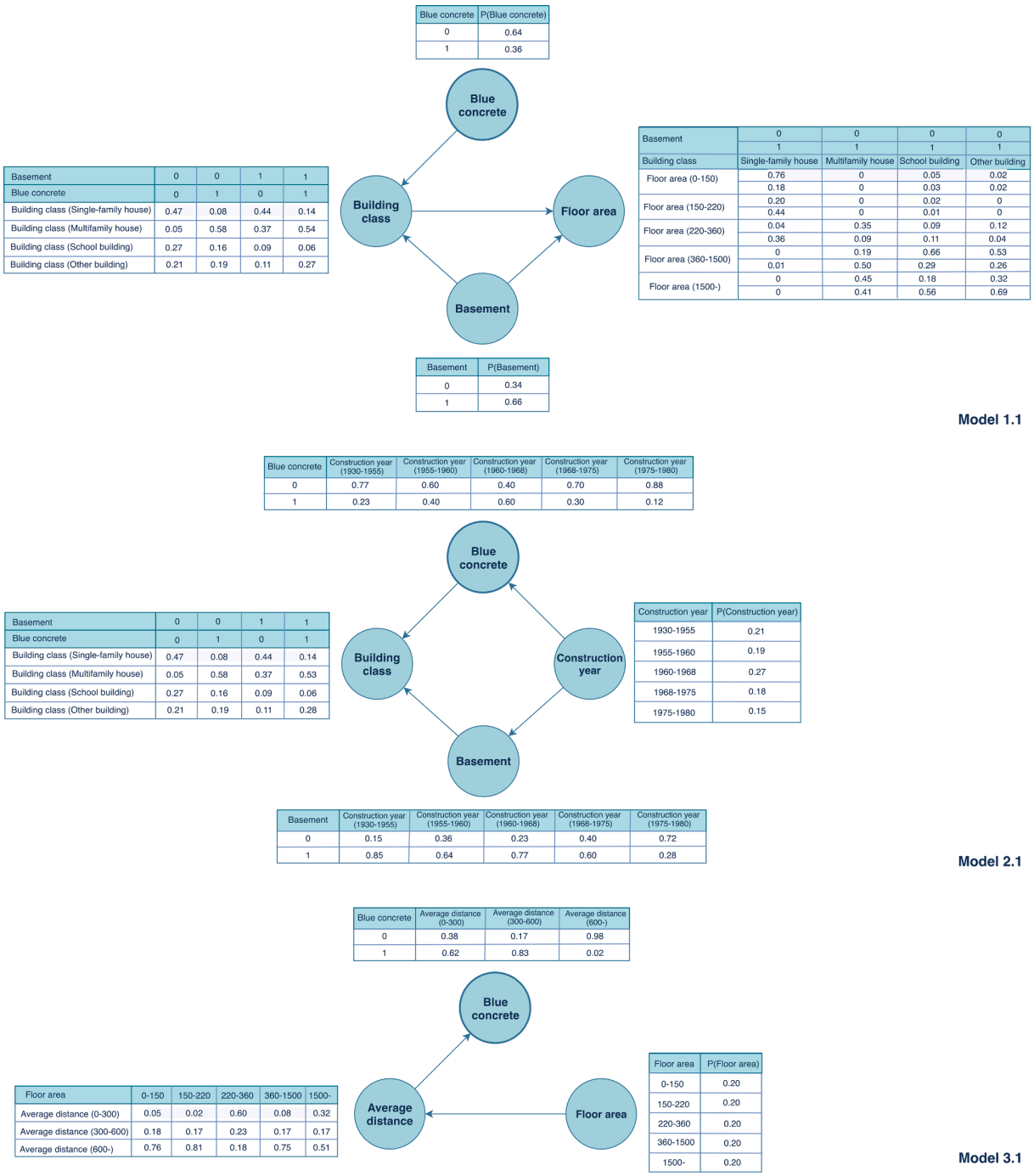


Fig. 6. Bayesian networks for the selected models.

Then the *Bayesian information criteria* (*BIC)* was employed as the scoring function to evaluate the individual model's performance and the average performance for the given input node sets. High individual scores signify the fit between models and the underlying data patterns, whereas average scores indicate critical node sets. The BIC metric balances the model complexity, i.e., the number of nodes, and the prediction performance, meanwhile preventing the risk of overfitting. The results of the three top models, with the highest average and individual scores, from model training, are described in Appendix C and summarized below:

- Model 1.1:

P(Floor area, Building class, Basement, Blue concrete) = Pr (Building class | Blue concrete) Pr (Floor area | Building class) Pr (Floor area | Basement) Pr (Building class | Basement).

- Model 2.1:

P(Construction year, Building class, Basement, Blue concrete) = Pr (Building class | Blue concrete) Pr (Basement | Construction year) Pr (Blue concrete | Construction year) Pr (Building class | Basement).

- Model 3.1:

P(Floor area, Average distance, Blue concrete) = Pr (Blue concrete |

Table 4
Local BIC scores were calculated using various variables as blue concrete parent nodes.

Variable importance	Conditional dependency	Score	Difference
Baseline	Pr (*Blue concrete* | None)	−359	0
1	Pr (*Blue concrete* | *Average distance*)	−160	199
2	Pr (*Blue concrete* | *Average distance, Building class*)	−162	197
3	Pr (*Blue concrete* | *Average distance, Construction year*)	−180	179
4	Pr (*Blue concrete* | *Average distance, Floor area*)	−194	165
5	Pr (*Blue concrete* | *Building class*)	−313	46
6	Pr (*Blue concrete* | *Number of apartments*)	−313	46
7	Pr (*Blue concrete* | *Floor area*)	−316	43
8	Pr (*Blue concrete* | *Average distance, Building class, Construction year*)	−318	41
9	Pr (*Blue concrete* | *Building class, Construction year*)	−329	30
10	Pr (*Blue concrete* | *Number of stairwells*)	−329	30

Table 5
Independencies between factors in the Bayesian network models.

Independencies	
Causal model 1.1	(Basement ⊥ Blue concrete)
	(Blue concrete ⊥ Basement)
	(Blue concrete ⊥ Floor area | Basement, Building class)
	(Floor area ⊥ Blue concrete | Basement, Building class)
Causal model 2.1	(Basement ⊥ Blue concrete | Construction year)
	(Blue concrete ⊥ Basement | Construction year)
	(Construction year ⊥ Building class | Basement, Blue concrete)
	(Building class ⊥ Construction year | Basement, Blue concrete)
Causal model 3.1	(Blue concrete ⊥ Floor area | Average distance)
	(Floor area ⊥ Blue concrete | Average distance)

Average distance) Pr (Average distance | Floor area).

The edges to these models were built subsequently as DAG skeletons to compute conditional probabilities distributions (CPDs) in parameter learning. Bayesian parameter estimator, a more conservative parameter learning using the existing prior CPDs and updating according to pseudo state counts before normalization, was adopted to model the networks' probabilistic relationships (or edge weights). The basic K2 prior and a more sensitive BDeu prior (Bayesian Dirichlet equivalent uniform prior) were tested to assess the variance between CPDs. Then the CPDs were related to the DAG skeletons for each model to construct Bayesian networks, presented in Fig. 6.

To summarize, the likelihood of detecting blue concrete is estimated to be 36% among the observed buildings. Based on the CPDs in Model 1.1, the probabilities of containing blue concrete in multifamily houses (53–58%) and other buildings (19–28%) are much higher than in single-family houses or school buildings, regardless of the existence of basements. Model 2.1 further shows that roughly 60% of buildings built between 1960 and 1968 are more likely to contain blue concrete than in other construction periods. The least probability of detecting blue concrete was observed in buildings constructed between 1975 and 1980 (12%). Also, buildings situated over 600 km away from blue concrete plants are found rarely contain blue concrete, according to Model 3.1. The results are reasonable considering the coherency between the historical blue concrete timeline and their detection in residential buildings in literature. The period of 1960–1975 corresponds to the Million Programme in Sweden, when large numbers of public housing were built in a short time, the same decades during which blue concrete was frequently used in construction. The geographical differences in the blue concrete detection likelihood also explain the local production and usage in nearby regions.

3.2.2. Network analytics

In search of edge patterns in various DAGs, the models were evaluated with incrementing evidence, and the top-scoring results from the predictive analysis are summarized in Appendix C. Scenario I was constructed based on the key evidence identified in models 1.1 and 2.1, including construction year, floor area, building class, and blue concrete. Then additional factors, the average distance from model 3.1, were appended to Scenario II together with basements and the number of floors. Scenario III is the most comprehensive with all the evidence, including the number of stairwells and apartments. The finding shows that models' performance improves progressively with the increase of evidence; in this case, Scenario III is preferred. From the constructed DAGs, dependencies were observed recurrently between blue concrete and construction year or average distance, as well as building class and floor area. The results are coherent with the correlation matrixes in Fig. 5, and thus the developed networks are considered valid.

Furthermore, the diagnosis analysis identified dominant nodes to blue concrete by computing BIC scores under different combinations of parent nodes to ascertain the degree of changes in the posterior probability distribution, presented in Table 4. The findings show that "average distance to blue concrete manufacturing plants" is the most significant factor, followed by combined factors of "average distance and building class", "average distance and construction year", and "average distance and floor area". For the individual factor, "number of apartments" and "number of stairwells" are also somehow influential in blue concrete detection. Again, the outcomes agree with the results in the previous predictive analysis. Building class, ranking the second critical individual factor in the diagnosis analysis, data should be stratified accordingly in modeling Bayesian networks for blue concrete.

The last part of network analytics deals with model robustness, where the extent to which conditional dependencies are affected by changes in the dataset structure is evaluated. The sensitivity analysis was performed on the subsets partitioning by building classes and variable rebinning. The networks with the highest score are illustrated in Appendix D. Overall, the bayesian network for school buildings fits the data subset better than other building classes with a higher BIC score. The average distance to blue concrete manufacturing plants is found to be a common contributing factor to the presence of blue concrete across building classes. The optimal DAGs tailored for each building class are presented below:

$P_{\text{Single-family house}}$ (Construction year, Average distance, Basement, Blue concrete) = Pr (Blue concrete | Average distance) Pr (Basement | Average distance) Pr (Construction year | Basement).

$P_{\text{Multifamily house}}$ (Floor area, Average distance, Blue concrete) = Pr (Average distance | Blue concrete) Pr (Floor area | Average distance).

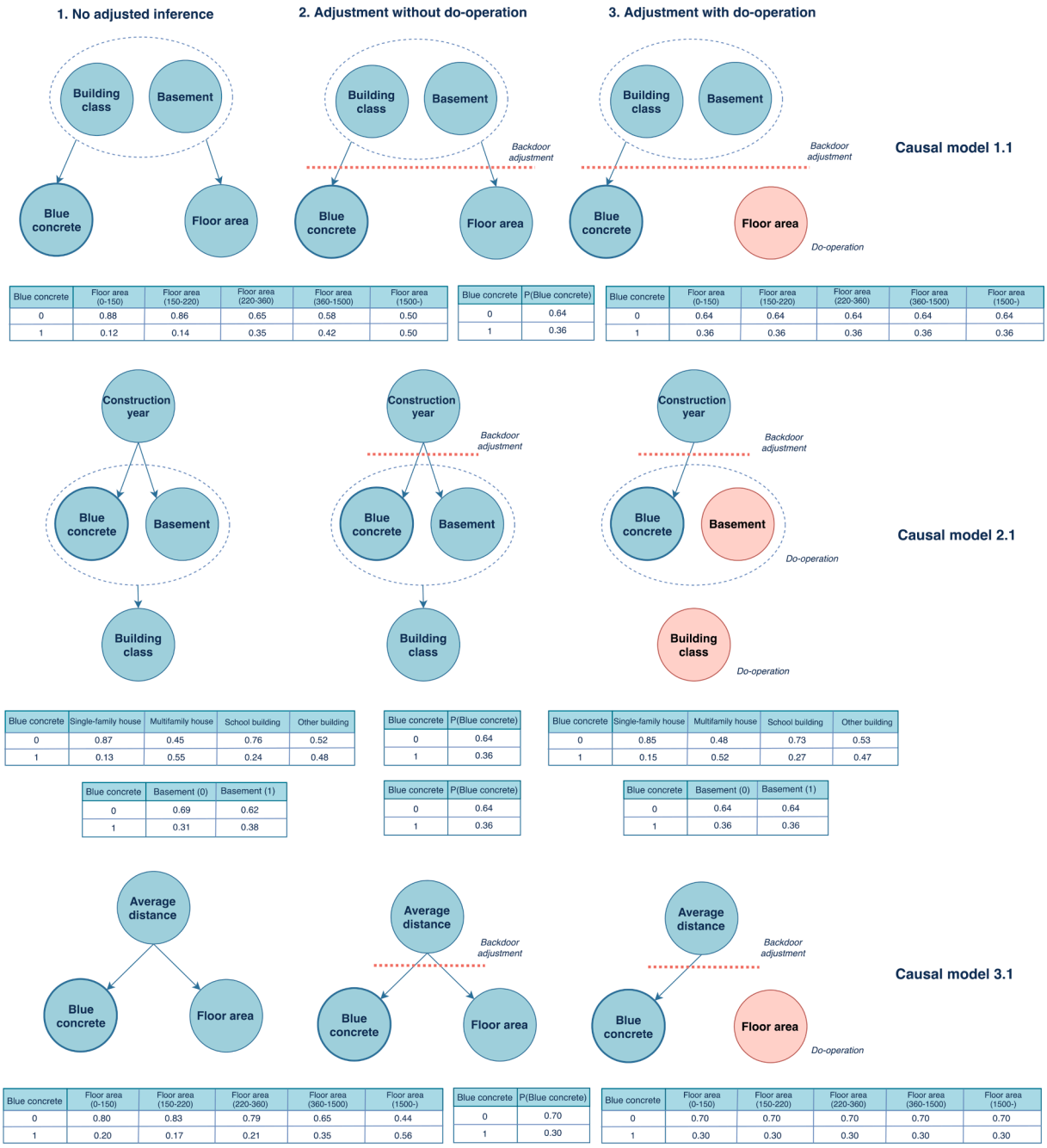
$P_{\text{School building}}$ (*Floor area, Average distance, Blue concrete*) = Pr (*Average distance* | *Blue concrete*) Pr (*Floor area* | *Blue concrete*).

$P_{\text{Other building}}$ (*Construction year, Floor area, Average distance, Number of stairwells, Blue concrete*) = Pr (*Average distance* | *Blue concrete*) Pr (*Number of stairwells* | *Average distance*) Pr (*Floor area* | *Number of stairwells*) Pr (*Construction year* | *Floor area*).

3.2.3. Causal inference

Prior to transforming Model 1.1–3.1 into causal network models, the factor independencies were computed, presented in Table 5. Causal model 1.1 shows that blue concrete is independent of the basement and floor area, but dependent on the combined factors of basement and building class. Causal model 2.1 aligns with the partial results from Causal model 1.1 that blue concrete is independent of the basement, but dependent on the construction year. Construction year, on the other hand, is independent of building class but rather dependent on the combined factors of the basement and blue concrete. Causal model 3.1 also agrees with the results from causal models 1.1 and 2.1 and indicates that blue concrete is independent of floor area but dependent on the average distance to the historical blue concrete manufacturing plants.

Furthermore, adjustment sets and do-operations were performed on



Blue concrete	Floor area (0-150)	Floor area (150-220)	Floor area (220-360)	Floor area (360-1500)	Floor area (1500-)
0	0.88	0.86	0.65	0.58	0.50
1	0.12	0.14	0.35	0.42	0.50

Blue concrete	P(Blue concrete)
0	0.64
1	0.36

Blue concrete	Floor area (0-150)	Floor area (150-220)	Floor area (220-360)	Floor area (360-1500)	Floor area (1500-)
0	0.64	0.64	0.64	0.64	0.64
1	0.36	0.36	0.36	0.36	0.36

Blue concrete	Single-family house	Multifamily house	School building	Other building
0	0.87	0.45	0.76	0.52
1	0.13	0.55	0.24	0.48

Blue concrete	P(Blue concrete)
0	0.64
1	0.36

Blue concrete	Single-family house	Multifamily house	School building	Other building
0	0.85	0.48	0.73	0.53
1	0.15	0.52	0.27	0.47

Blue concrete	Basement (0)	Basement (1)
0	0.69	0.62
1	0.31	0.38

Blue concrete	P(Blue concrete)
0	0.64
1	0.36

Blue concrete	Basement (0)	Basement (1)
0	0.64	0.64
1	0.36	0.36

Blue concrete	Floor area (0-150)	Floor area (150-220)	Floor area (220-360)	Floor area (360-1500)	Floor area (1500-)
0	0.80	0.83	0.79	0.65	0.44
1	0.20	0.17	0.21	0.35	0.56

Blue concrete	P(Blue concrete)
0	0.70
1	0.30

Blue concrete	Floor area (0-150)	Floor area (150-220)	Floor area (220-360)	Floor area (360-1500)	Floor area (1500-)
0	0.70	0.70	0.70	0.70	0.70
1	0.30	0.30	0.30	0.30	0.30

Fig. 7. Causal Bayesian network models 1.1–3.1.

the causal models. The existence of active front door and backdoors paths in each model was ascertained to determine possible adjustment sets for path blocking. Afterward, do-operation was conducted on the adjusted inference and translated observational distributions (probabilities without the do-operator) into interventional distribution (probabilities with the do-operator). Fig. 7 displays causal inference results for different models and the probability distributions of blue concrete in non-adjustment, adjustment without do-operation, and adjustment with do-operation. The findings show that floor area, in causal models 1.1 and 3.1, and basement, in causal model 2.1, are not the parameters indicating the occurrence of blue concrete. However, building class is pinpointed as the indicator for blue concrete based on the change of posterior probabilities of the target variables.

3.3. Probabilistic inference in the regional building database

Around 115,996 buildings built between 1930 and 1980 from the regional building database were employed as input data to the trained network models. The approximate inference was queried based on the Bayesian hierarchical models with the evidence of blue concrete for different variables as examples. The inference results of individual models and the model averages are illustrated in Table 6, of which the joint probabilistic distribution of the queried variables sums to 100% and the disjoint probabilistic distribution of each evidence state also adds to 100%. Overall, blue concrete is estimated to be present in 33.7% of the regional building stock. Multifamily houses had the highest risk of containing blue concrete than other buildings and single-family houses, while school buildings are ranked the lowest risk. It is estimated that 19.7% of multifamily houses, 8.6% of other buildings, 4.1% of single-

Table 6
Bayesian hierarchical modeling for the probabilistic distributions of blue concrete in the regional building stock conditioned on different variables: (6a) Blue concrete, (6b) Building class, (6c) Construction year, (6d) Basement, (6e) Floor area, (6f) Average distance to blue concrete plants.

Table 6**a.** Probabilistic distribution of blue concrete.

Blue concrete	Model 1.1	Model 2.1	Model 3.1	Average
0	63.0%	64.9%	70.9%	66.3%
1	37.0%	35.1%	29.1%	33.7%

Table 6**b.** Probabilistic distribution of building class.

Joint distribution		Model 1.1		Model 2.1		Model 3.1		Average	
Building class	Evidence	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint
Single-family house	Blue concrete (0)	28.3%	44.6%	28.8%	45.0%	–		28.6%	44.8%
	Blue concrete (1)	4.1%	12.1%	4.1%	12.1%	–		4.1%	12.1%
Multifamily house	Blue concrete (0)	16.7%	26.7%	16.7%	25.1%	–		16.7%	25.9%
	Blue concrete (1)	19.7%	54.1%	19.6%	54.5%	–		19.7%	54.3%
School building	Blue concrete (0)	9.7%	14.5%	9.6%	15.1%	–		9.7%	14.8%
	Blue concrete (1)	3.5%	9.3%	3.3%	9.0%	–		3.4%	9.2%
Other building	Blue concrete (0)	9.0%	14.1%	9.6%	14.8%	–		9.3%	14.5%
	Blue concrete (1)	9.0%	24.5%	8.2%	24.4%	–		8.6%	24.5%

Table 6**c.** Probabilistic distribution of construction year.

Joint distribution		Model 1.1		Model 2.1		Model 3.1		Average	
Construction year	Evidence	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint
1930–1955	Blue concrete (0)	–		16.3%	25.4%	–		16.3%	25.4%
	Blue concrete (1)	–		4.8%	14.0%	–		4.8%	14.0%
1955–1960	Blue concrete (0)	–		11.6%	17.3%	–		11.6%	17.3%
	Blue concrete (1)	–		7.6%	20.7%	–		7.6%	20.7%
1960–1968	Blue concrete (0)	–		10.8%	17.1%	–		10.8%	17.1%
	Blue concrete (1)	–		15.6%	45.5%	–		15.6%	45.5%
1968–1975	Blue concrete (0)	–		12.4%	20.3%	–		12.4%	20.3%
	Blue concrete (1)	–		5.7%	14.5%	–		5.7%	14.5%
1975–1980	Blue concrete (0)	–		13.3%	20.0%	–		13.3%	20.0%
	Blue concrete (1)	–		2.0%	5.3%	–		2.0%	5.3%

Table 6**d.** Probabilistic distribution of Basement.

Joint distribution		Model 1.1		Model 2.1		Model 3.1		Average	
Basement	Evidence	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint
0	Blue concrete (0)	22.8%	33.9%	22.9%	36.9%	–	22.9%	35.4%	
	Blue concrete (1)	11.8%	34.4%	11.6%	30.0%	–	11.7%	32.2%	
1	Blue concrete (0)	41.8%	66.1%	40.7%	63.1%	–	41.3%	64.4%	
	Blue concrete (1)	23.7%	65.6%	24.8%	70.0%	–	24.3%	67.8%	

Table 6**e.** Probabilistic distribution of floor area.

Joint distribution		Model 1.1		Model 2.1		Model 3.1		Average	
Floor area	Evidence	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint
0–150	Blue concrete (0)	11.2%	18.5%	–		15.1%	22.6%	13.2%	20.6%
	Blue concrete (1)	1.6%	4.6%	–		3.5%	13.5%	2.6%	9.1%
150–220	Blue concrete (0)	10.1%	15.8%	–		17.2%	23.3%	13.7%	19.6%
	Blue concrete (1)	1.8%	4.8%	–		3.6%	11.6%	2.7%	8.2%
220–360	Blue concrete (0)	11.0%	16.7%	–		16.2%	22.9%	13.6%	19.8%
	Blue concrete (1)	5.9%	15.3%	–		4.1%	13.4%	5.0%	14.4%
360–1500	Blue concrete (0)	16.6%	26.5%	–		13.5%	18.8%	15.1%	22.7%
	Blue concrete (1)	12.2%	34.0%	–		6.9%	24.3%	9.6%	29.2%
1500-	Blue concrete (0)	14.8%	22.6%	–		9.2%	12.5%	12.0%	17.6%
	Blue concrete (1)	14.8%	41.5%	–		10.8%	37.2%	12.8%	39.4%

Table 6**f** Probabilistic distribution of average distance to historical blue concrete manufacturing plants.

Joint distribution		Model 1.1		Model 2.1		Model 3.1		Average	
Average distance	Evidence	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint	Joint	Disjoint
0–300	Blue concrete (0)	–		–		8.0%	11.7%	8.0%	11.7%
	Blue concrete (1)	–		–		13.8%	44.2%	13.8%	44.2%
300–600	Blue concrete (0)	–		–		3.1%	4.5%	3.1%	4.5%
	Blue concrete (1)	–		–		15.1%	51.7%	15.1%	51.7%
600-	Blue concrete (0)	–		–		58.9%	83.8%	58.9%	83.8%
	Blue concrete (1)	–		–		1.2%	4.1%	1.2%	4.1%

family houses, and 3.4% of school buildings in the regional buildings were built with blue concrete. In terms of construction year, buildings built between 1960 and 1968 are most likely to be detected with blue concrete, followed by 1955–1960. On the other hand, buildings constructed between 1930 and 1955 or 1968–1975 are less likely to contain blue concrete, and the least risky group is the period 1975–1980. The joint distribution of the Bayesian hierarchical model also indicates that 15.6% of buildings built during 1960–1968, 7.6% of

Table A1
Overview of the blue concrete dataset.

Value category	Data specification	Measurement type
Building-generic data from the national building dataset		
1. Matching keys	National real estate index	String + Nominal
	EPC index (Energy declaration index)	Nominal
	FNR (Real estate key)	Nominal
	UUID (Universally unique identifier)	String
	Address	
2. Cadastral info	Municipality	String
	Postcode	Nominal [5 digits]
	Post place	String
3. Building usage	Municipality building category code	Nominal [1–7 types]
	Municipality building usage code	Nominal [1–99 types]
	EPC building category	Nominal [Single or double-family house, Multifamily house, Non-residential building]
	EPC building type	Nominal [Detached, semi-attached, attached]
4. Building characteristics	Building age	Scale variable [Year]
	Floor area	Scale [m^2]
	Number of floors	Ordinal
	Number of stairwells	Nominal
	Number of apartments	Scale
	Number of basements	Nominal [0, 1, 2, >2]
	Shelter room	Binary [Yes, No]
5. Ventilation	Ventilation type	Nominal [Exhaust, Balanced, Balanced with heat exchanger, Exhaust with heat pump, Natural ventilation]
6. Radon	Indoor radon annual average value	Scales [Bq/m^3]
Building-specific data from municipality indoor radon measurements		
7. Blue concrete	Blue concrete detection	Binary [Yes, No]
5. Building characteristics	Foundation type - Gävle	Nominal [Basements, Suspended foundation, Souterrain, Shallow foundation, Unknown]
6. Radon	Indoor radon annual average value	Scales [Bq/m^3]
Building-specific data from pre-demolition audit inventories		
7. Blue concrete	Blue concrete detection	Binary [Yes, No]
	Blue concrete-containing component	Nominal [Wall, facade, floor/foundation, ventilation shaft, others, unspecific]

buildings built during 1955–1960, 5.7% of buildings built during 1968–1975, 4.8% of buildings built during 1930–1955, as well as 2.0% of buildings built during 1975–1980 potentially contain blue concrete.

Concerning building sizes, larger buildings have a higher risk of containing blue concrete. For instance, buildings with a floor area between 360–1500 m^2 have almost doubled the risk of blue concrete than those under 360 m^2. Approximately 12.8% of buildings with a size above 1500 m^2 and 9.6% of buildings between 360–1500 m^2 are suspected of containing blue concrete, while only 5.0% of median and around 2,5% of small buildings potentially have blue concrete. Notably, buildings built with at least one basement have a doubled probability of having blue concrete compared to those built without, where 24.3% of the buildings with basements and 11.7% of the buildings without basements are probably exposed to blue concrete. Finally, the average distance to blue concrete manufacturing plants was proved to be a determining factor. Buildings situated further than 600 km have a significantly low risk of contamination by blue concrete.

4. Discussion

The section discussed the key findings and compared the applicability of the Bayesian network method with prediction approaches. The last part highlights the practical implementation of the developed models for in situ hazardous building material assessment, including blue concrete.

4.1. Results implication

The data-driven pipeline for blue concrete pattern identification was configured and demonstrated using Bayesian network models. Approximately 18% of the observed buildings contain blue concrete in the blue concrete dataset, which matches the expert estimation of 15–20% blue concrete-containing residential dwellings and workplaces from national indoor radon measurements (Rönnqvist, 2021). In the study, the detection frequency and the link between blue concrete and indoor radon were characterized through detailed aggregation at the building class level than the current literature (Clavensjö & Åkerblom, 2020). The study outcomes contribute to knowledge expansion about blue concrete from residential to non-residential buildings and provide an overview of the radon situation in existing building stocks. The findings on the types of blue concrete elements are beneficial for appraising remediation actions and related implementation costs for blue concrete-induced radon in different building classes (Clavensjö & Åkerblom, 2020). Besides, the impacts of crucial cadastral and building parameters on radon concentration were determined and aligned with previous literature (McGrath & Byrne, 2020), assuring data validity and pinpointing critical features for subsequent network modeling.

Choosing multiple Bayesian networks with the highest BIC scores is favorable for model comparison and complementation of probability distributions for factors that are not modeled in specific networks. It is observed that joint probability distributions Pr (Yi | Xi) in Models 1.1–3.1 are mutable, depending on the numbers and the types of variables included in structural learning. Therefore, by solely identifying dependent relationships from data without involving domain knowledge, the direction of the edges between P (event | prior knowledge) can sometimes be misleading. For example, P (Building class | Blue concrete), with blue concrete as the parent node and building class as the children node, does not seem to be plausible to formulate a query of "what is the probability that the building class is a single-family house given the presence of blue concrete?". In fact, the question will be reasonable and queriable if the order of the variables is reversed, which signifies the necessity of employing domain knowledge for preliminary DAG calibration. However, the algorithms did capture the underlying patterns of the data, i.e., dependency or correlation between variables, and the links between nodes are stable across models. For instance, blue concrete is dependent of building class (Model 1.1, 1.2 and Scenario I), construction year (Model 1.2, Scenario I and II), the average distance to manufacturing plants (Model 3.1, Scenario II and III) but is independent of floor area (Model 1.1, 3.1, and Scenario I-III).

Further evaluating the results from network analytics, the edges in DAG are changeable in case of more variables are incorporated into network training. The strength of the dependent relationships between nodes variates along with the node dynamics. This can be seen in the predictive analysis, where the link between building class and blue concrete was replaced by the average distance to manufacturing plants in Scenario I and II, as well as the link between construction year blue concrete disappeared in Scenario III compared to Scenario I and II. The assumption of variable domination was verified in the diagnosis analysis. The average distance to blue concrete plants, by itself, is the most significant variable among all the others. This variable and its combination with building class, construction year, and floor area influence the presence of blue concrete. Besides, the findings from sensitivity analysis show that simple networks are favorable for single-family houses, multifamily houses, or school buildings. The exception is the

Table B1
Overview of the arithmetic mean and confidence intervals (CI) of the annual average of indoor radon levels by building classes, basements, and ventilation types (values in bold were computed with a minimum of 20 observations).

Building class	Radon CI	Blue concrete	Basement	Radon CI	Ventilation type	Radon CI
Single-family house (n = 1,831)	**130 ± 10**	No detection	Without	**70 ± 17**	Natural	**83 ± 25**
					Exhaust	44 ± 13
					Balanced	57 ± 13
			With	**76 ± 14**	Natural	**80 ± 16**
					Exhaust	50 ± 15
					Balanced	72 ± 27
		Detection	Without	187 ± 76	Natural	235 [-]
					Exhaust	177 ± 89
					Balanced	–
			With	**367 ± 140**	Natural	367 ± 173
					Exhaust	275 ± 113
					Balanced	671 [-]
Multifamily house (n = 263)	**167 ± 16**	No detection	Without	98 ± 30	Natural	110 [-]
					Exhaust	72 ± 13
					Balanced	156 ± 10
			With	**112 ± 16**	Natural	**125 ± 25**
					Exhaust	**104 ± 26**
					Balanced	**92 ± 35**
		Detection	Without	**247 ± 36**	Natural	**290 ± 37**
					Exhaust	129 ± 34
					Balanced	–
			With	**190 ± 35**	Natural	**235 ± 73**
					Exhaust	**176 ± 41**
					Balanced	128 ± 54
School building (n = 58)	**73 ± 13**	No detection	Without	**71 ± 16**	Natural	–
					Exhaust	6 [-]
					Balanced	**72 ± 16**
			With	**67 ± 16**	Natural	–
					Exhaust	–
					Balanced	67 ± 16
		Detection	Without	114 ± 16	Natural	–
					Exhaust	106 ± 0
					Balanced	130 [-]
			With	116 ± 67	Natural	–
					Exhaust	–
					Balanced	116 ± 68
Other building (n = 54)	**192 ± 75**	No detection	Without	**205 ± 85**	Natural	540 ± 0
					Exhaust	70 ± 54
					Balanced	**135 ± 78**
			With	66 ± 40	Natural	–
					Exhaust	240 [-]
					Balanced	49 ± 24
		Detection	Without	455 ± 167	Natural	426 ± 223
					Exhaust	–
					Balanced	540 [-]
			With	225 ± 225	Natural	–
					Exhaust	93 [-]
					Balanced	244 ± 244

model for other buildings whose DAG is complicated but linear. This could be due to the observations clustered in this subset being rather heterogeneous and hard to identify consistent patterns. The school buildings, on the other hand, have homogeneous building usage and typology that allows a simpler DAG structure and better fit.

The results from causal inference suggest that the presence of blue concrete depends on average distance and construction year, but the indirect impact from other covariates, such as building class and basements, should not be disregarded. Although blue concrete is independent of basements and floor area, the combined presence of blue concrete and basements can be found more frequently in certain building types. It would be thus necessary to stratify data subgroups according to their building classes in model training before implementing causal inference. Despite of generic modeling approach, the results of probability inference approximately aligns with the conditional probability distributions in parameter learning with a slight variance between models. Based on conditional probability in Model 1.1, around 36% of the buildings are likely to contain blue concrete. Using the identical DAG structure from Model 1.1 and performing variable binning with the same interval on the prediction dataset, 37% of regional buildings built between 1930 and 1980 with blue concrete were estimated in the probabilistic inference. The share of blue concrete from the samples and the large population is close, indicating the certainty of the predicted results. In the end, slight adjustments were made to compute averaging probability of 34% for a conservative estimation.

4.2. Comparison between predictive approaches

Learning Bayesian networks has several advantages for hazardous building material prediction in terms of operability and explainability. It combines the strengths of machine learning classification, which identifies data patterns without the need for explicit programming, and the interpretability of traditional statistic methods. Unlike conventional Bayesian models that require expert knowledge to specify variable dependency and probability distribution, learning the Bayesian network harnesses structure learning to generate potential DAG structures and then perform parameter learning based on the input dataset. Because of this feature, it is flexible and efficient to update new CPDs when more instances are added to the training dataset. Its inference outputs –causal graphical networks and comprehensive probabilistic lookup tables –

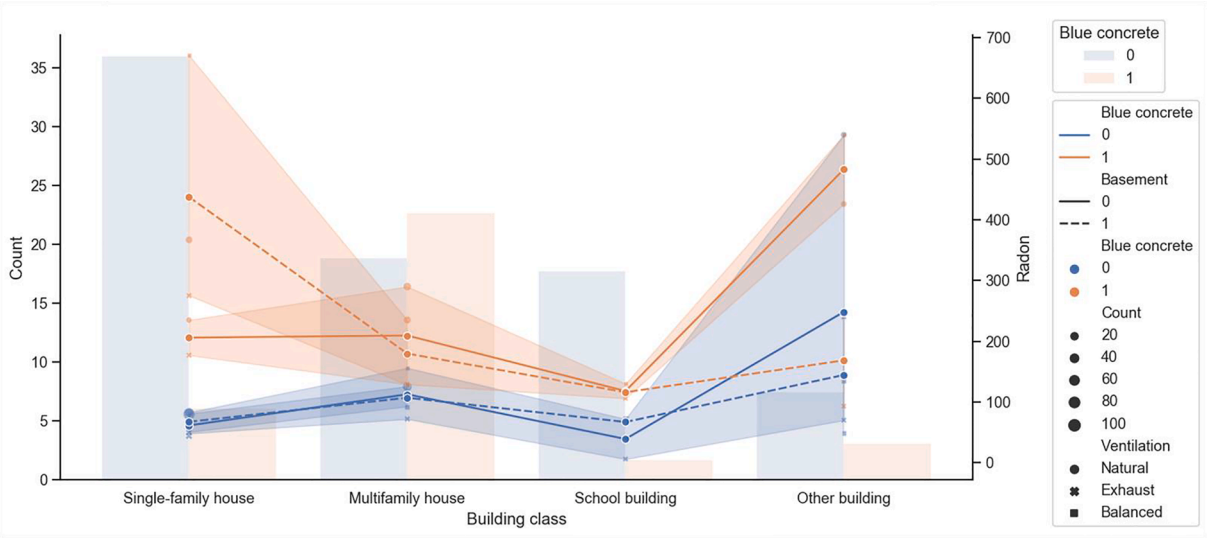


Fig. B1. Averaging radon level by building class, blue concrete, basement, and ventilation type illustrated in scatterplots (considering data size of each specific data subgroup) and bar plots (counts of blue concrete and non-blue concrete subgroups by building class). The overall results were summarized in line charts with arithmetic mean values and confidence intervals (CI) of annual indoor radon levels.

Table C1
Exploring potential DAG of the blue concrete Bayesian networks using structural learning algorithms and evaluated with the BIC scoring metric. The three highest scores of the node sets were sorted in descending order, and the best models with the highest scores are marked in bold.

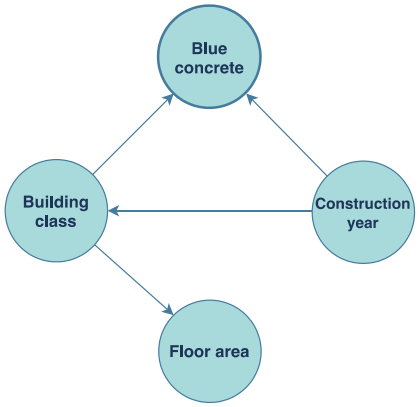
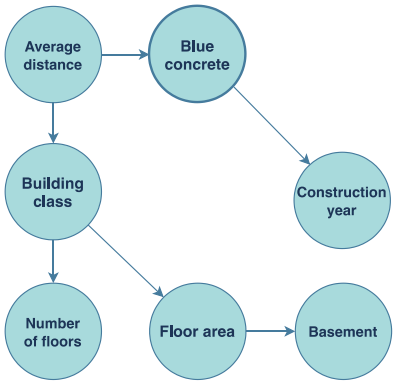
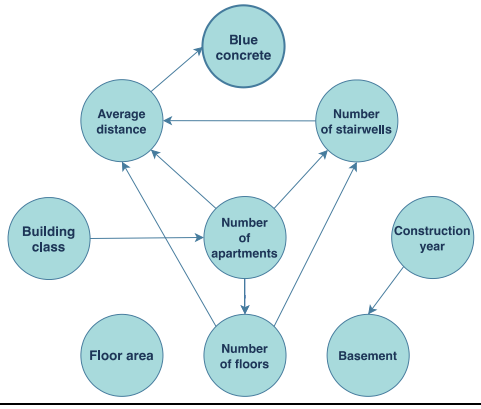
Rank		Input	Hyperparameter	Scoring method		Structural learning
No	Model	Nodes	Algorithm	BIC	Average	Edges
1	1.2	Floor area, Building class, Basement,	hc	−2731	−2739	(Blue concrete, Building class), (Building class, Floor area), (Basement, Floor area), (Basement, Building class)
	1.4	Blueconcrete	ts	−2754		(Blue concrete, Building class), (Building class, Floor area), (Floor area, Basement)
	1.1		**pc**	**−2729**		**(Blue concrete, Building class), (Building class, Floor area), (Basement, Floor area), (Basement, Building class)**
	1.3		mmhc	−2742		(Building class, Blue concrete), (Floor area, Building class), (Floor area, Basement), (Basement, Building class)
2	2.3	Construction year, Building class, Basement,	hc	−3162	−3121	(Blue concrete, Building class), (Construction year, Blue concrete), (Construction year, Building class), (Basement, Construction year), (Basement, Building class)
	2.2	Blueconcrete	ts	−3098		(Blue concrete, Building class), (Building class, Construction year), (Construction year, Basement)
	2.4		pc	−3174		(Blue concrete, Construction year), (Blue concrete, Building class), (Construction year, Building class), (Basement, Construction year), (Basement, Building class)
	2.1		**mmhc**	**−3051**		**(Blue concrete, Building class), (Construction year, Basement), (Construction year, Blue concrete), (Basement, Building class)**
3	3.2	Floor area, Average distance, Blue concrete	hc	−3153	−3138	(Blue concrete, Average distance), (Blue concrete, Floor area), (Average distance, Floor area)
	3.1		**ts**	**−3123**		**(Blue concrete, Average distance), (Average distance, Floor area)**
	3.2		pc	−3153		(Average distance, Floor area), (Blue concrete, Floor area), (Blue concrete, Average distance)
	3.1		**mmhc**	**−3123**		**(Floor area, Average distance), (Average distance, Blue concrete)**

aree rather intuitive and transparent compared to other gray-box or black-box prediction approaches. In addition, learning Bayesian networks can handle small datasets and missing values; meanwhile, it does not need tedious feature engineering and model tuning. The characteristics make it particularly suitable for modeling unstructured and heterogeneous building-specific data.

However, applications of learning Bayesian networks in the building sector are rare. Only a few studies are found in the areas of building inspection and diagnosis (Pereira et al., 2021), building predictive maintenance (Bortolini & Forcada, 2017), building performance evaluation (Bortolini & Forcada, 2020), and building damage prediction in disasters (Chen & Zhang, 2021). These studies reported that Bayesian networks are robust and rigorous for quantifying multivariate probability under uncertainty and, thus, can be swiftly replicated for other building elements providing necessary adaptation. However, no Bayesian network use cases are found for hazardous building material prediction in the literature. It may be due to the fact that building stock is extremely complicated with various levels of systems and materials. Extending the scope of a single Bayesian network model for comprehensive inference may not be as effective as integrating several Bayesian models. The overwhelming models' complexity and limited possibility of validation may be the reasons restricting the feasible implementation of Bayesian networks to real problems.

Nonetheless, learning Bayesian networks have some drawbacks primarily related to accuracy. Among all, the probabilistic inference comes with a lower granularity due to the need for data discretization. The

Table D1
The evolvement of the DAGs in relation to the incrementing evidence.

Hyperparameter		Metric		Structural learning
Algorithm	**Scoring method**	**Score**	**Average**	**DAG construction**
Scenario I				
P (*Construction year, Floor area, Building class, Blueconcrete*) = Pr (*Blueconcrete* | *Construction year*) Pr (*Blueconcrete* | *Building class*) Pr (*Building class* | *Construction year*) Pr (*Floor area* | *Building class*)				
pc / mmhc	BDeu	−5091	−5080	
	K2	−5033		
	BIC	−5117		
Scenario II				
P (*Construction year, Floor area, Building class, Average distance, Basement, Number of floors, Blueconcrete*) = Pr (*Building class* | *Average distance*) Pr (*Floor area* | *Building class*) Pr (*Number of Floors* | *Building class*) Pr (*Construction year* | *Blue concrete*) Pr (*Blue concrete* | *Average distance*) Pr (*Basement* | *Floor area*)				
ts / pc	BDeu	−4756/ −4759	−4771	
	K2	−4728/ −4736		
	BIC	−4729/ −4718		
Scenario III				
P (*Construction year, Floor area, Building class, Average distance, Basement, Number of floors, Number of stairwells, Number of apartments, Blueconcrete*) = Pr (*Average distance* | *Number of floors*) Pr (*Number of stairwells* | *Number of floors*) Pr (*Average distance* | *Number of apartments*) Pr (*Number of floors* | *Number of apartments*) Pr (*Number of apartments* | *Building class*) Pr (*Number of stairwells* | *Number of apartments*) Pr (*Average distance* | *Number of stairwells*) Pr (*Blue concrete* | *Average distance*) Pr (*Basement* | *Construction year*)				
pc	BDeu	−3862	−4031	
	K2	−3799		
	BIC	−4431		

trade-off derives from the fact that too large variable intervals will result in a rough model, whereas too fine intervals will lead to few unrepresentable data points. Besides, some searching algorithms, such as Exhaust Search and Max-Min Hillclimb Search, are computationally expensive, thus, are not able to handle high dimensional data. This, in turn, limits the model capability in the search for optimal model parameters and hyperparameters. Therefore, prior knowledge is needed in preliminary node selection to delineate the search scope. Machine learning does not have such limits and is able to cope with high dimensional, massive, and mixed types of input data. But at the same time, they are sensitive to missing values and require a relatively large dataset to avoid overfitting. In spite of excellent prediction performance, the prediction results from machine learning are unable to discover the causal relationships between variables (Wu et al., 2022). Consequently, learning Bayesian networks have its edge in hazardous building material prediction for preliminary screening and approximate reasoning.

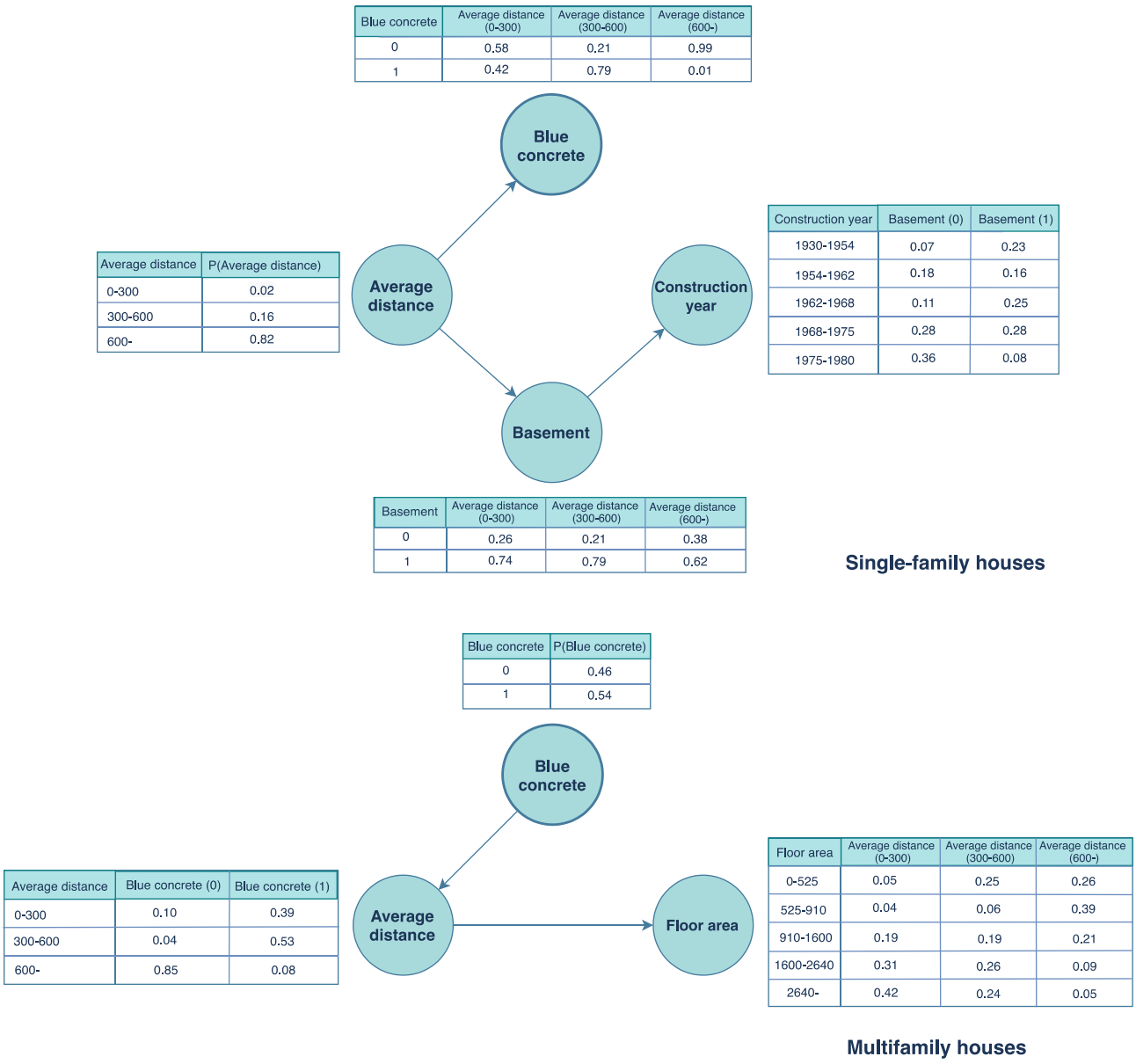


Fig. E1. Bayesian networks trained on the specific building class subsets: (1) Single-family houses, (2) Multifamily houses, (3) School buildings, (4) Other buildings.

4.3. *Practical implementation*

The data-driven Bayesian approach can advise risk-based inspections when prioritizing the building groups with higher probabilities of containing in situ hazardous materials. The outcomes, despite primary, provide additional information for building auditors for blue concrete assessment during the early inspection procedure. It is also relatively cost-effective compared to the previous radiation scanning with the vehicle for mapping the presence of blue concrete buildings. The Bayesian network model is also reproducible when introducing new observations to the data pool and updating posterior probabilities of the target variables accordingly, which is especially beneficial in modeling a dynamic urban environment. Nowadays, there are roughly 300 properties built with blue concrete in Umeå municipality, according to vehicle radiation measurements from early times (Umeå Kommun, 2020). With the constant development of the city, this number may be outdated, and hard to trace the material flows of blue concrete in renovation or demolition activities. The predictive inference method developed in this study is able to overcome the limitation and implement it on various scales for purposes. On the one hand, the models can be used to identify risk-prone building groups and devise tailored policies for relevant authorities. On the other hand, the query can also be made at individual buildings to evaluate the likelihood of encountering blue concrete for property owners or demolition contractors.

To scale up the probabilistic inference from the dataset to the regional or national scale, data representativeness was controlled to minimize potential sampling and selection bias. The representativeness of the observed buildings was addressed by comparing the building class distribution of the regional and national building stocks. The constitution of the sample dataset consists of 88% residential and 12% non-residential buildings. The proportions resemble the building registers from five municipalities in 2021, where residential dwellings count for 92%, and the rest of the non-residential buildings are 8% (Statistics Sweden, 2021). Further examining the national building stock, the proportions of building types are also similar −94% residential and 6% non-residential dwellings (Statistics Sweden, 2021). To further improve the models' generalizability and the results' granularity, adding new observations from the five municipalities or other municipalities to the

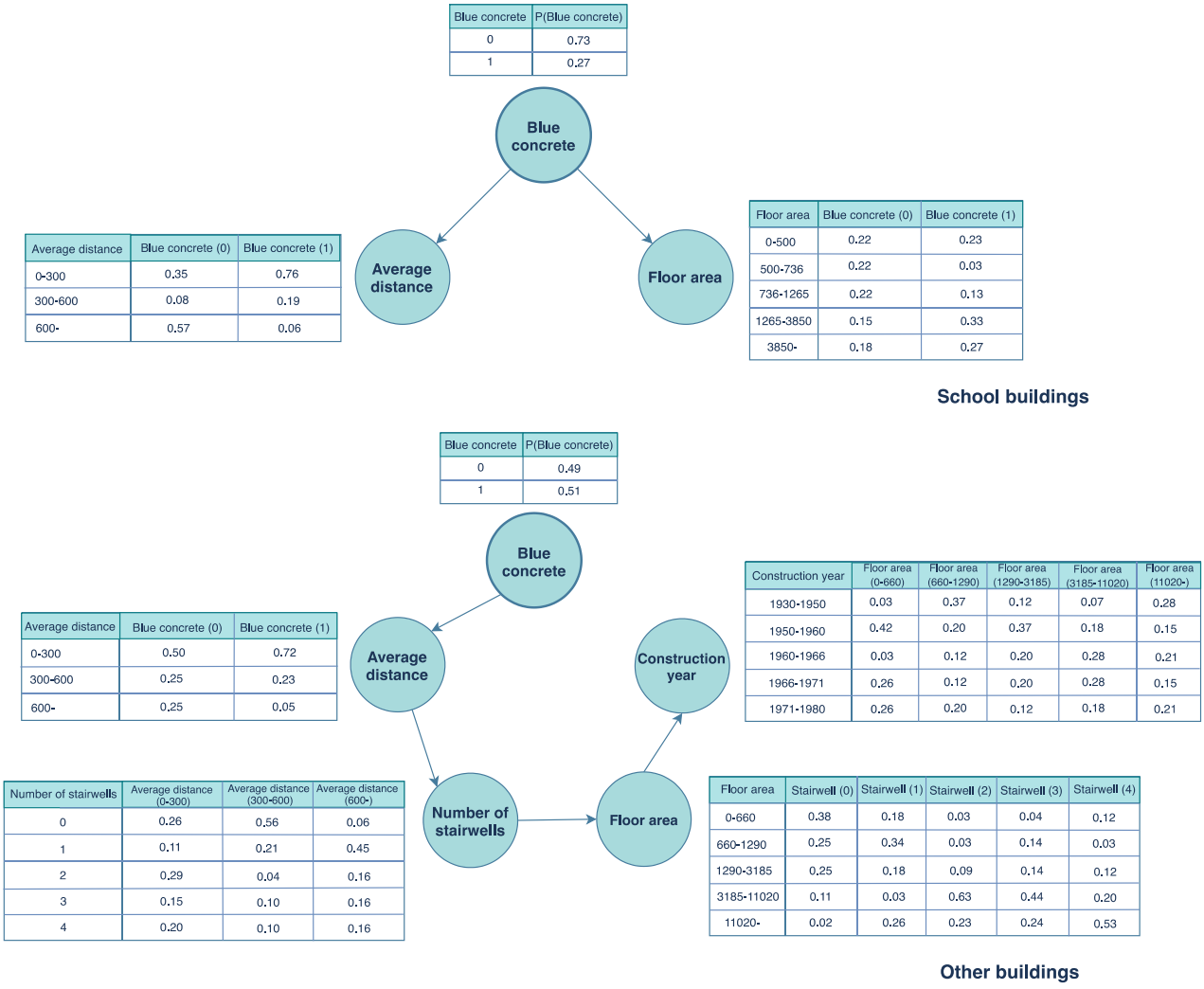


Fig. E1. (*continued*).

existing training dataset is needed. The refined inference can be used to resolve the confounding effect on indoor radon from blue concrete and ground sources.

5. Conclusions

The study investigates the possibility of identifying the patterns of residual blue concrete through predictive inferences based on inspection records, indoor radon measurements, and building registers. Blue concrete is estimated to be present in approximately 34% of buildings built between 1930 and 1980, more common than the existing assumptions. Training learning Bayesian networks on the input data enable one to untangle independencies and compute conditional probabilities between blue concrete and other variables. The findings show that the average distance to blue concrete manufacturing plants is the most critical attribute for inferring the presence of blue concrete, followed by building class and construction year. Basements and floor area are independent of the occurrence of blue concrete in the causal inference. By further applying the developed models to the registers of uninvestigated buildings in the sampled municipalities, the risk-prone building groups with higher probabilities of containing blue concrete are highlighted.

To the authors' best knowledge, it is the first study developing a learning Bayesian networks pipeline for hazardous building material prediction in building stock. The proposed predictive approach is reproducible by updating probabilistic inferences by adding new samples. The prediction outcomes could guide risk-based inspections to evaluate buildings with potential blue concrete contamination and form a basis for radon remediation planning. The primary limitations of the study are the variety and sufficiency of the training data, which restricts a more detailed inference for particular building classes. Future research is suggested to include more blue concrete inspection records from other municipalities to improve models' generalizability and validate the developed models empirically.

Author contributions

P.-Y.W. collected and compiled pre-demolition audit inventories, conducted data preprocessing and model building, and wrote and revised the manuscript. T.J. assembled indoor radon measurements, merged them with their registered data, and created a blue concrete dataset. M.M., T.J., and C.S. participated in the method discussion and manuscript review. K.M. acquired the research fund and reviewed the

manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

The research fund comes from the Swedish Foundation for Strategic Research (SSF) with grant number FID18-0021, the Re:Source project from the Swedish Energy Agency with grant number P2022-00304, and the EU BuiltHub project with grant agreement ID of 957026.

CRediT authorship contribution statement

Pei-Yu Wu: Conceptualization, Methodology, Formal analysis, Validation, Writing – original draft, Writing – review & editing, Visualization. **Tim Johansson:** Supervision, Investigation, Data curation, Writing – review & editing. **Mikael Mangold:** Supervision, Funding acquisition, Writing – review & editing, Project administration. **Claes Sandels:** Supervision, Writing – review & editing. **Kristina Mjörnell:** Supervision, Funding acquisition, Writing – review & editing, Project administration.

Declaration of Competing Interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

The work is part of the PhD project "Prediction of Hazardous Materials in Buildings using Machine Learning" supported by RISE Research Institutes of Sweden. Special thanks are sent to Cecilia Jelinek from the Geological Survey of Sweden (SGU), who provided information on the radiation measurements with vehicles in the Swedish municipalities.

Appendix A. Data sources

Table A1

Appendix B. Characterization of blue concrete in buildings – Hierarchical clustering

Table B1
Fig. B1

Appendix C. Learning Bayesian networks for blue concrete – Structural learning

Table C1

Appendix D. Network analytics – Bayesian networks from the predictive analysis

Table D1

Appendix E. Network analytics – Bayesian networks from the sensitivity analysis

Fig. E1

References

A nkan, 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.

Barr, I. (2018). Causal Inference With Python Part 2 - Causal Graphical Models. Retrieved October 3, 2022, from http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/.

Björk, C., Kallstenius, P., & Reppen, L. (2013). Så byggdes husen 1880–2000: Arkitektur, konstruktion och material i våra flerbostadshus under 120 år. *Svenskbyggtjänst.*

Borgelt, C., & Kruse, R. (2001). An empirical investigation of the K2 metric. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 2143*, 240–251. https://doi.org/10.1007/3-540-44652-4_22

Bortolini, R., & Forcada, N. (2017). Discussion about the use of bayesian networks models for making predictive maintenance. *Decisions, 86*(July), 973–980. https://doi.org/10.24928/jc3-2017/0145

Bortolini, R., & Forcada, N. (2020). A probabilistic performance evaluation for buildings and constructed assets. *Building Research and Information, 48*(8), 838–855. https://doi.org/10.1080/09613218.2019.1704208

Bouabdallaoui, Y., Lafhaj, Z., Yim, P., Ducoulombier, L., & Bennadji, B. (2021). Predictive maintenance in building facilities: A machine learning-based approach. *Sensors (Switzerland), 21*(4), 1–15. https://doi.org/10.3390/s21041044

Boverket. (2013). *Åtgärder mot radon i bostäder*.

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

Chen, W., & Zhang, L. (2021). Predicting building damages in mega-disasters under uncertainty: An improved Bayesian network learning approach. *Sustainable Cities and Society, 66*(December 2020), 102689. https://doi.org/10.1016/j.scs.2020.102689.

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

DIGG. (2022). Sveriges dataportal. Retrieved January 27, 2022, from Myndigheten för Digital Förvaltning website: https://www.dataportal.se/sv.

Gao, X. G., Gao, Z. G., Ren, H., Yang, Y., Chen, Qing, D., & He, C. C. (2019). Learning Bayesian network parameters via minimax algorithm. *International Journal of Approximate Reasoning, 108*, 62–75. https://doi.org/10.1016/J.IJAR.2019.03.001

Hall, T., & Vidén, S. (2005, July). The million homes programme: A review of the great Swedish planning project. *Planning Perspectives*, Vol. 20, pp. 301–328. Taylor & Francis Group. https://doi.org/10.1080/02665430500130233.

Jelinek, C., & Eliasson, T. (2015). *Strålning från bergmaterial.*

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

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

Liu, X., Lu, Y., Xia, Z., Li, F., & Zhang, T. (2017). A data mining method for potential fire hazard analysis of urban buildings based on Bayesian network. *ACM International Conference Proceeding Series, Part, F1318*, 1–6. https://doi.org/10.1145/3144789.3144811

Liu, Z., Malone, B., & Yuan, C. (2012). Empirical evaluation of scoring functions for Bayesian network model selection. *BMC Bioinformatics, 13 Suppl 1*(Suppl 15). https://doi.org/10.1186/1471-2105-13-S15-S14

Lucchi, E. (2016). Multidisciplinary risk-based analysis for supporting the decision making process on conservation, energy efficiency, and human comfort in museum buildings. *Journal of Cultural Heritage, 22*, 1079–1089. https://doi.org/10.1016/J.CULHER.2016.06.001

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.

McGrath, J. A., & Byrne, M. A. (2020). An approach to predicting indoor radon concentration based on depressurisation measurements. *Indoor and Built Environment*, 1–9. https://doi.org/10.1177/1420326X20924747

Pearl, J., & Dechter, R. (2013). Identifying independencies in causal graphs with feedbacks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence* (p. 4).

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

QuantumBlack. (2020). Causal Inference with Bayesian Networks. Main Concepts and Methods — causalnex 0.11.0 documentation. Retrieved October 12, 2022, from https://causalnex.readthedocs.io/en/latest/04_user_guide/04_user_guide.html.

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

Sedin, D., & Hjelte, I. (2004). *The Radon Situation in Sweden.* (July), 3–5.

Statens offentliga utredningar från Näringsdepartementet. (2001). *Radonläget i Sverige*.

Statens offentliga utredningar, J. (1983). *Radon i bostäder*.

Statistics Sweden. (2021). Number of dwellings by region and type of building (including special housing). Year 2013 - 2021. Retrieved October 13, 2022, from Statistics

Sweden website: https://www.statistikdatabasen.scb.se/pxweb/en/ssd/START_BO_BO0104_BO0104D/BO0104T01/.

Swedish National Board of Housing Building and Planning. (2010a). *Radon in the indoor environment (Radon i inomhusmiljö)*. Retrieved from https://www.folkhalsomyndigheten.se/livsvillkor-levnadsvanor/miljohalsa-och-halsoskydd/inomhusmiljo-allmanna-lokaler-och-platser/radon/.

Swedish National Board of Housing Building and Planning. (2010b). *Technical status in Swedish buildings - results from the BETSI project (Teknisk status i den svenska bebyggelsen - resultat från projektet BETSI)*.

Swedish Radiation Safety Authority. (2013). *Measurement of radon in residential buildings - method description (Mätning av radon i bostäder - metodbeskrivning)*.

Taskesen, E. (2022). Causation — bnlearn bnlearn documentation. Retrieved August 29, 2022, from https://erdogant.github.io/bnlearn/pages/html/Structure learning.html.

Umeå Kommun. (2020). Blåbetong. Retrieved September 19, 2022, from https://www.umea.se/byggaboochmiljo/avfallochatervinning/materialsorteringvidrivningochrenovering/allamaterial/blabetong.4.4ff54ec174f999469c348.html.

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

Wu, P.-Y., Mjörnell, K., Mangold, M., Sandels, C., & Johansson, T. (2021). A data-driven approach to assess the risk of encountering hazardous materials in the building stock based on environmental inventories. *Sustainability (Switzerland), 13*(7836), 1–26.

Wu, P.-Y., Sandels, C., Mjörnell, K., Mangold, M., & Johansson, T. (2022). Predicting the presence of hazardous materials in buildings using machine learning. *Building and Environment, 213*(February). https://doi.org/10.1016/j.buildenv.2022.108894