Probabilistic modelling of building stock properties for urban mining

Lombe Mutale, Ramon Hingorani, Nils Dittrich and Jochen Köhler

Norwegian University of Science and Technology, Trondheim, Norway

Contact: lombe.mutale@ntnu.no

Abstract

The construction industry is one of the biggest contributors to greenhouse gas emissions and unsustainable waste. A circular economy of the existing building stock can contribute to minimising mining of finite resources and reducing the construction industry's waste. However, stakeholders often list lack of information about the existing building stock as a barrier against implementing a circular economy in the construction industry. This study provides a framework for construction industry stakeholders to combine publicly available data sources to obtain probability-based information about the building stock. The study analyses existing building data at city level using Bayesian Networks, a probabilistic modelling approach that accounts for the missing data consistently in contrast to other methods. The framework can be extended to incorporate first principle, data-based and empirical models from disciplines such as structural engineering, architecture, and industrial ecology to facilitate a circular economy.

Keywords: circular economy; probabilistic modelling; existing building stock; residential buildings.

1 Introduction

The built environment stock and construction industry are among the biggest contributors to global greenhouse gas emissions, energy consumption, and unsustainable waste. This can partly be attributed to the production of building materials and components, construction, and demolition. In fact, in Europe, one third of all waste stems from construction and demolition activities [1].

A circular economy of the existing building stock can contribute to minimising mining of finite resources and reducing the construction industry's waste. Citywide circular economy approaches aim at implementing this approach at the local level. This makes it easier to set up a framework for a circular economy when working together with relevant actors, but also reduces the need for transport that can counteract the economic and environmental benefits of reusing and recycling. However, stakeholders often list a lack of information about the existing building stock as a barrier against implementing a circular economy in the construction industry [2].

There are attempts to collate information about existing buildings from several European countries including footprint, height, building type, and age [3]. However, not only are there many missing entries for these attributes, but also, countries such as Norway were not included and important information like the main construction material is missing. Moreover, there is no suggestion on the part of the researchers on how to deal with data scarcity.

Probabilistic modelling is one way to effectively capture information on the building stock to facilitate circularity, allowing us to deal with an increased level of uncertainty due to either random

chance, or lack of knowledge [4]. An application can be found in [5], where information from different data sources was collected and merged to make inferences based on geospatial, material, and typologies building data. In contrast to the current study where the uncertainty of the data is part of the model, they use statistical approximations to fill in missing data fields.

For this reason, we suggest Bayesian Networks (BNs) - a probabilistic modelling tool characterised by directed acyclic graphs (DAGs) of nodes representing variables with causal (dependent) relationships i.e., directed links without loops or cycles. Each node has a conditional probability table (CPT) with a probability for each possible state, given the state of its parents. The prior CPTs can be created directly from data or reflect expert information. Thereafter, evidence is set in one or more of the nodes resulting in the updating of the entire network's CPTs [6]. Not only have BNs and causal inference been used in the medical field and several others but also for structural reliability, design optimisation, and risk assessment in civil engineering [7, 8].

BNs provide the opportunity to combine data or observations with scientific principles. This makes it possible to expand data sets with experts' opinions and prior scientific findings without explicitly requiring the data that these are based on. Compared to other popular machine learning approaches, BNs display some more key advantages. These include a probabilistic nature, meaning all results are explicitly uncertain, overall good performance, and high interpretability. The latter characteristics can be summed up under the term "grey box", which stands in between highly interpretable but poorly performing models (e.g., linear regression) or highly performing but difficult to interpret models (e.g., neural networks).

In this paper we present a case study to apply BNs to building stock information from the City of Trondheim in the northern part of south Norway in the Trøndelag region. The city has a population of about 200 000 people and 77 641 buildings of which 40 167 of them are residential buildings mostly constructed after 1960 [9].

We gathered information about a sample of existing buildings from national databases. We then used this information to create a probabilistic model with BNs. Finally, we made inferences from the model about other existing buildings for which we do not have data. The purpose of the case study was to illustrate the benefits and possibilities of using probabilistic modelling and BNs, but it is not exhaustive and can be built upon in the future.

The study methodology is described in Section 2. The results and discussion are presented in Section 3 and conclusions in Section 4.

2 Methodology

The following section presents the study methodology – refer to Figure 1. We first gathered building stock variables from publicly available databases. Next, we compared and analysed the variables to test the fidelity of the data available in each database. This then determined which database and variables to select and use for the study. Finally, we used these variables to create a BN for probabilistic modelling and material prediction.

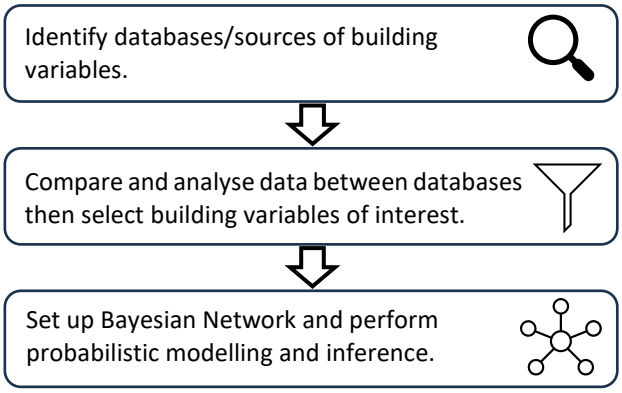

Figure 1. Study methodology.

2.1 Database identification and data gathering

The study focusses on four building variables – main construction material (Material), year of construction (Year), Type, and total floor area (Area). The building variables were sourced from two databases: Geodata Norway and City Antiques databases referred to as Geodata and Antiques respectively in this paper.

Geodata is a geographical information systems (GIS) platform with, among others, downloadable

files about buildings [10]. The Trondheim buildings' geodatabase contains information on 75 634 buildings (Table 1) i.e., almost all the buildings in Trondheim.

Table 1. Number of buildings for each variable in each database and missing data (X).

Variable	Data type	Geodata	Antiques
No. buildings	-	**75634**	**7313**
Year	Interval	62203	5043
Type	Nominal	74752	**X**
Material	Nominal	**X**	4547
No. stories	Interval	74252	**X**
Footprint area	Ratio	75634	**X**

In addition, Trondheim Municipality has an ongoing project of collecting and digitising historical building information from various sources. From this project, the Antiques department has compiled a geographical database that comprises data for 7313 buildings (Table 1)[11] that have been selected for conservation. The two databases are linked by building numbers i.e., the unique identifier assigned to each building. This is useful because it allows us to combine data from both databases thereby increasing the information we have (Table 1).

2.2 Variable description

There are a total of 7313 buildings which are in both the Geodata and Antiques databases. Although there is an overlap of building variables in the databases, there is also inhomogeneity and scarcity of data. This is further exemplified in Table 2 which shows that on the one hand, the Geodata database has Year, Type, No. of stories, and Footprint area available. On the other hand, the Antiques database has Year and Material available but not Type, No. of stories, and Footprint area. Additionally, Footprint area is the only variable we have on all buildings while there is information missing for all the other variables.

Table 2. Number of buildings in one or both databases for each variable and missing data (X).

Variable	Geodata	Antiques
No. of buildings	**7313**	**7313**
Year	4024	5043
Type	7147	**X**
Material	**X**	4547
No. stories	6967	**X**
Footprint area	7313	**X**

2.2.1 Year

The construction year is available in both the Antiques and Geodata databases. The Year is listed for 4024 and 5043 buildings in the Geodata and Antiques databases respectively as shown in Table 2 and Figure 2.

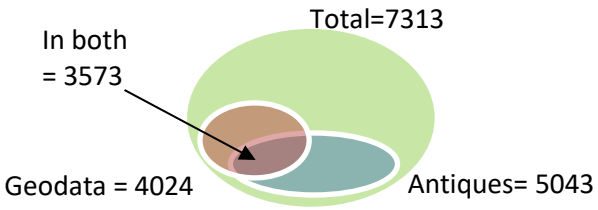

Figure 2. Venn diagram describing the number of buildings in the sample with valid entries for the property Year in each database.

We elected to use the "Year" variable from the Antiques database since it has a higher number of entries than the Geodata database.

2.2.2 Type

The building's use type found in the Geodata database is based on the Statistics Norway's three level classification system [12]. The Geodata database provides a different description of buildings than the Antiques database. The Antiques database's building type describes what can be viewed as a subset of the Statistics Norway type. Therefore, we elected to use the type in the Geodata database in this study.

2.2.3 Material

The Antiques database provides the main building material for 4547 buildings. However, as an example of missing data, there are 2766 buildings whose main material is listed as unknown. The

building materials listed are Timber, Masonry (Brick), Concrete, Stone (Natural Stone) and Steel.

2.2.4 Area

The shape area or footprint of the building is available in the Geodata database as is the number of building stories. Area is a derived property which was included as the product of the Footprint area and No. of building stories. Area is assumed to approximate the total floor area of each building. Neither Geodata nor Antiques database have the area and height for each building story.

2.2.5 Discretization of building data

The selected variables were divided into the categories listed in Table 3.

Table 3. Building variables and categories.

Year	Type	Material	Area
A: Before 1920	A: Rowhouses	A: Timber	A: Less than 249sqm
B: 1920 to 1944	B: Detached houses	B: Masonry	B: 250 to 499sqm
C: 1945 to 1969	C: Semi-detached houses	C: Concrete	C: 500 to 749sqm
D: 1970 to 1999	D: Large houses	D: Stone	D: 750 to 999sqm
E: 2000 and after	E: Holiday-Shared-Farm house	E: Steel	E: 1000sqm or more

The construction year categories were grouped to align with a study on the European building stock [13] as well as release years of the Norwegian Regulations on Technical Requirements for Construction Works (TEK) which are likely to have influenced building practices.

The material categories were selected based on the number of materials listed in the databases. Nonetheless, it is worth noting that there was no description of the material percentage in a building to qualify for the category. For instance, there are several composite buildings (e.g. steel and reinforced concrete) in Norway as identified by other researchers [14].

Residential buildings or dwellings make up more than half of the Trondheim building stock [9]. Therefore, it was decided to focus only on residential buildings, a common strategy for building stock studies [15–17].

Statistics Norway have categorised residential buildings as rowhouses, detached houses, semi-detached houses, large houses (multi-dwelling buildings), holiday houses, shared houses and farmhouses [12]. The last 3 are less common in the database and were therefore combined into a single category as shown in Table 3.

Area categories were selected to have a middle category of 500 to 749 square meters (sqm) based on sample descriptive statistics that showed that the median was ≈450 sqm and the 75 percentile of the data is ≈730 sqm. Two categories on either side of the middle category were added in keeping with five categories as for the other variables.

2.3 Setting up the Bayesian Network

2.3.1 Setting up the Directed Acyclic Graph

The Directed Acyclic Graph (DAG) was determined using a combination of automated structure learning and engineering judgement. For the former, we used Probabilistic Graphical Models Python (PGMPY)'s ranked exhaustive search algorithm [18] to find the most optimal relationships between the data. This resulted in several ranked DAGs. The chosen DAG used to create the BN is illustrated in Figure 3.

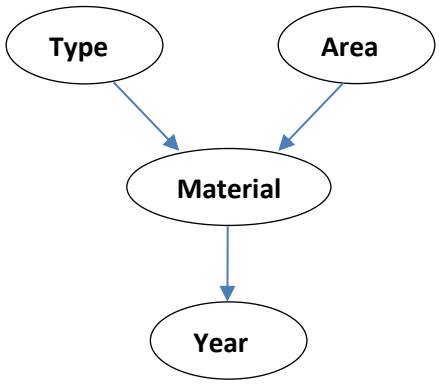

Figure 3. Bayesian Network with parent nodes of Material being Type and Area, and Year as child node of Material.

The structural equations for the Bayesian Network are presented in Eqn. 1 and 2.

$$Material := f(Type, Area) \quad (1)$$

where Type, and Area are causes or parents of Material.

$$Year := f(Material) \quad (2)$$

where Material is the cause or parent of Year.

We used the PGMPY package [19] to perform the analysis of the BN due to its availability and ease of use. Structural learning was performed using the Bayesian Estimator rather than the Maximum Likelihood estimator to avoid underestimation due to several cases of missing data.

2.3.2 Calculation of prior probabilities

The prior shows the probability before any conditioning and inference. In effect, prior probabilities represent the knowledge that we already have before performing inferences on the data. For each category i, its relative frequency or prior probability is its number of observations n_i, divided by the sum of observations of all categories, as presented in Eqn. (3).

$$p(x_i) = \frac{n_i}{\sum_i^n n_i} \quad (3)$$

For instance, there are 5043 observations with the Year classification (the others have no entry for the Year variable) – refer to Table 2. Of those observations, 2671 of them are in the category 'A: Before 1920' and therefore the probability for this category is 0,530. Note that the sum of probabilities for categories must equal 1.

Model validation was confirmed by checking that prior probabilities calculated by PGMPY resulted in the same numbers as a manual calculation.

2.3.3 Calculation of posterior probabilities

For inference or posterior probabilities, the joint probability mass function of the BN is calculated based on the chain rule for DAGs as presented in Eqn. 4 [6]:

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p[(x_i \,|pa(x_i)] \quad (4)$$

where the BN has *n* variables, and the right-hand side refers to the product of the probability of each variable, x_i conditioned on its parents, pa(x_i) in the DAG.

Therefore, the joint probability mass function of the BN (refer to Figure 3) is:

$$\begin{aligned} &p(Material, Area, Year, Type) \\ &= p(Material|Type, Area) \cdot p(Type) \\ &\cdot p(Area) \cdot p(Year\,|Material) \end{aligned} \quad (4)$$

Evidence is applied to one or more nodes to infer information from the BN. The probability of one node, given another node as evidence, can be determined using Eqn. 5.

$$p(x_i|\, x_j) = \frac{p(x_i\,,\, x_j)}{p(x_j)} \quad (5)$$

where variable, x_i is conditioned on variable x_j.

3 Results and discussion

Based on the number of buildings per database and per variable in Table 2, Year and Material are two of the building variables with the scarcest data. Therefore, the following section will focus on these two variables and how they relate to the other variables.

3.1 Probability of Type given the Year

In this section, we use the BN to investigate the probability of finding a certain Type when given the Year.

Figure 4's "Prior" row shows that types with categories A to D are almost evenly distributed with probabilities of 0,25, 0,27, 0,21 and 0,22 respectively. In contrast, Statistics Norway lists Type B: detached houses as making up around half of all residential buildings in Trondheim [20]. Nonetheless, they estimate that Type C: Semi-detached and Type A: Rowhouses have a frequency of 0,21 and 0,22 respectively, similar to what is found here. Contrastingly, Type D: Large dwellings (multi-dwelling buildings) have a 0,08 frequency.

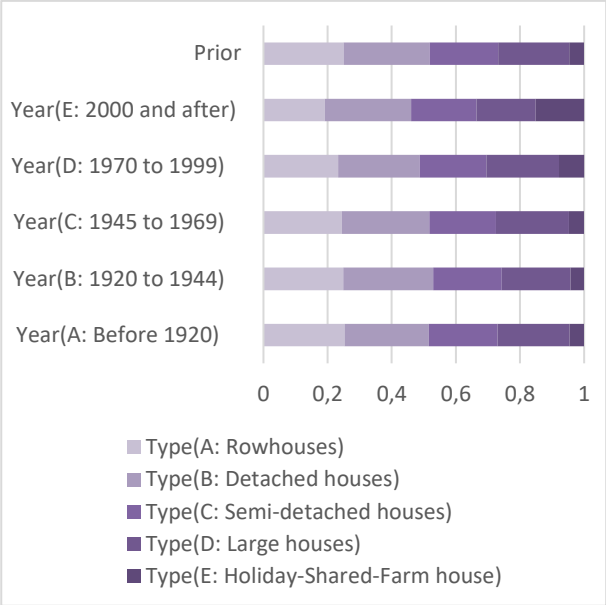

Figure 4. Conditional probability of building Type with varying Year - P(Type | Year).

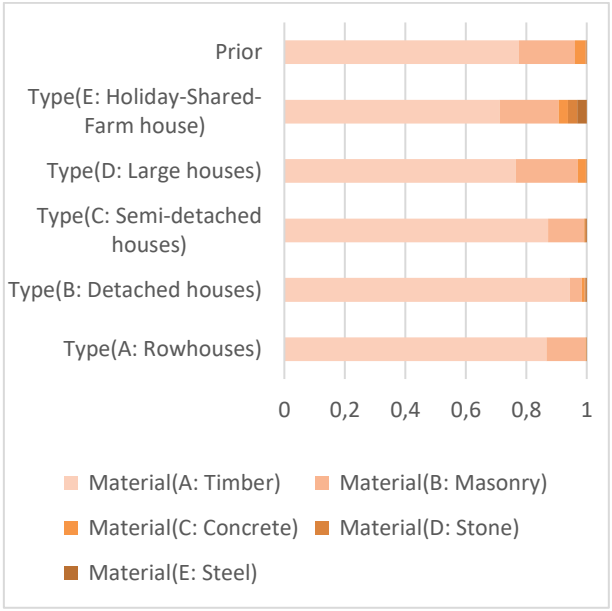

Figure 5. The conditional probability of building Material with varying Type - P(Material| Type).

Applying evidence i.e., conditioning on Year, reveals that the probability of Type A, B, and C does not change significantly depending on the Year. For example, the probability of Type B: Detached dwellings for Years A to E respectively are 0,26, 0,28, 0,27, 0,25, and 0,27 respectively.

Nevertheless, there is an increase in Type D: Large houses and Type E: Holiday-Shared-Farm houses with time. Their combined probabilities are 0,26, 0,26, 0,28, 0,30, and 0,34 for Years A to E respectively. This suggests that the fraction of apartment buildings and shared houses are increasing with time.

3.2 Probability of Material given the Type

This section presents the results of inferring the Material given the Type.

As illustrated in Figure 5 in the "Prior" row, 77% of residential buildings in the sample were constructed with timber. A fifth were constructed with masonry, and 4% with concrete. Less than one percent are made from stone and steel materials. These results are consistent with Norway's history of building with timber. As of the year 2000, 74% of all existing Norwegian dwellings were constructed from timber [21].

Even if we condition on the Type, Timber remains the most common building material with over 0,7 probability for all types. Still, there is a clear difference in probability depending on the residential building type.

For instance, the probability of Type B: Detached houses being constructed from timber increases from the prior of 0,77 to a posterior of 0,94. In contrast, the probability of Type E: Holiday-Shared-Family houses being constructed from timber decreases from the prior of 0,77 to a posterior of 0,71.

3.3 Probability of Year given the Material

Figure 6 shows the building year category probabilities for Material categories prior to inference and after inference.

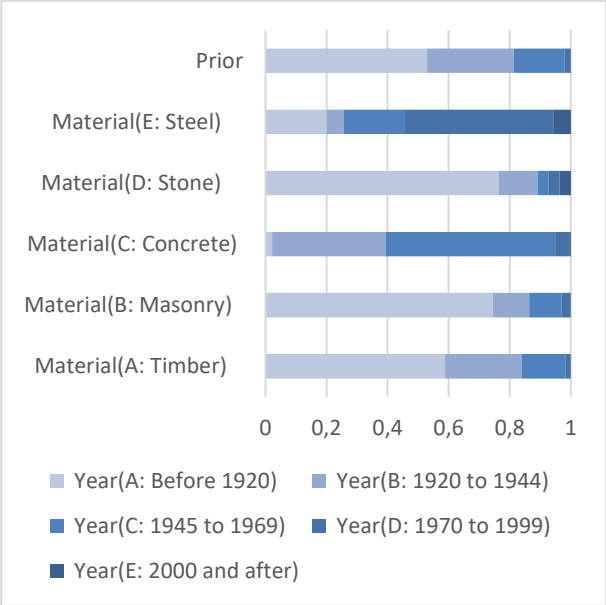

Figure 6. Conditional probability of building Year with varying Material - P(Year| Material).

The prior probabilities show that 53% of the buildings in the sample were constructed before 1920 and 28% were built between 1920 and 1944. 17% of the buildings were constructed between 1945 and 1969. Therefore, close to 98% of the buildings are over 53 years old. The aged population is to be expected considering that the sample is for buildings that have been earmarked for conservation.

Conditioning on the Material node reveals that Steel residential buildings have a 0,49 probability of being constructed in Year D: 1970 and 1999. The 0,49 probability for that period is higher than the probability of the preceding and succeeding Year categories. Similarly, concrete residential buildings have 0,56 and 0,37 probability of being constructed in Year C: 1945 to 1969 and Year B: 1920 to 1944 respectively. In contrast, material categories, Timber, Masonry, and Stone have a probability of being constructed in Year A: Before 1920 of 0,59, 0,74, and 0,76. This would suggest that these more natural materials became less common after 1920 in favour of steel and concrete.

3.4 Probability of Area given the Type and Material

If we were to further use the information that we have gained so far and condition the building Area on both Type and Material = Timber, we obtain the results shown in Figure 7.

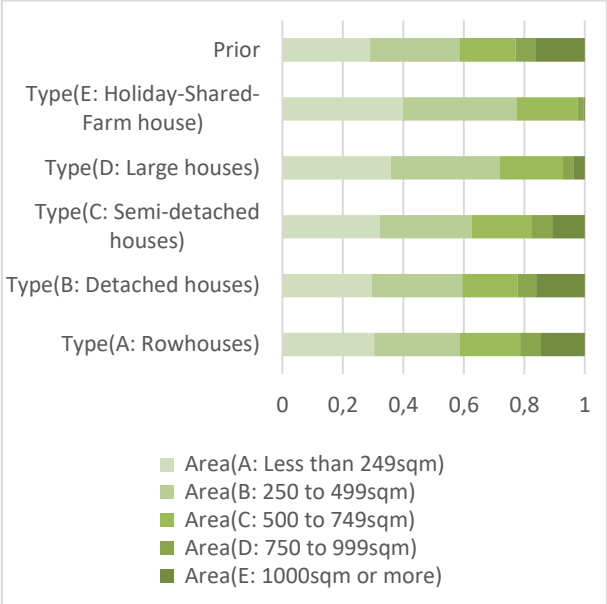

Figure 7. Conditional probability of building Area with varying Material - P(Area| Type, Material=A: Timber).

The prior results show that 77% of residential buildings have total areas less than 750 square meters (based on combining probabilities of categories Area: A, B, and C).

The conditioned results reveal that area probabilities change with type of timber residential building. The combined probability of Area: A, B, and C is 0,79, 0,78, 0,83, 0,93, and 0,98 for Type A, B, C, D, and E respectively. This result makes sense considering that the types are related to the size of the building.

3.5 General discussion

3.5.1 Brief overview

A sample of buildings in Trondheim, and four building variables Year (of construction), Material, Area, and Type was selected for preliminary analysis of the methodology. These variables were sourced from two databases and used to create a BN for probabilistic modelling and prediction of unknown quantities in the Trondheim building stock.

The results show that a simple BN can be used for bottom-up estimates of building stock properties.

Despite the simplicity of the model, the results make sense and agree with other literature and statistics.

3.5.2 Critical analysis of findings

The results presented here are preliminary and simply a means to test out the methodology. Other factors such as the neighbourhood, building height, renovation history etc., will also influence the results. Moreover, it is important to consider the limitations in the sample of buildings before applying it to the population. For instance, the results here are specifically for Trondheim residential buildings constructed before 2007.

3.5.3 Implications

The case study can be extended to include more variables in the network and georeferenced as done by other researchers [3]. Thereafter, the results can be used for e.g., estimations of reusable components, material intensity or materials estimations for calculating embodied carbon [15].

4 Conclusions

In order to promote a circular use of building components and materials, we need information about what is currently available. Multiple national sources and databases provide information about existing buildings. However, the data is often inhomogeneous, in some cases missing or incomplete. We suggest BNs to account for the missing and scarce data in a probabilistic way.

The next step would be to find more data sources and extend the network by increasing the number of variables. Additionally, rather than categories, we could assign random variables and distributions for the prior data and conditioning observations. This will lead to a better understanding of the limits and uncertainties of the data and the observations which can then be applied to the population i.e., Trondheim building stock. Moreover, there can be further investigation into structure learning methods and incorporation of discipline-specific knowledge such as structural engineering, industrial ecology and architecture to determine the best network.

Finally, such a BN could be part of a dynamic platform that is connected to different databases and updated continuously. In this way, probabilistic information can be available for construction industry stakeholders to make sustainable and informed decisions.

5 Acknowledgements

This paper is part of the Circular City project, funded by the Norwegian University of Science and Technology through the SusRes program.

6 References

[1] European Commission. Construction and demolition waste, https://environment.ec.europa.eu/topics/waste-and-recycling/construction-and-demolition-waste_en (2023, accessed 9 January 2023).

[2] Tingley DD, Cooper S, Cullen J. Understanding and overcoming the barriers to structural steel reuse, a UK perspective. *Journal of Cleaner Production* 2017; 148: 642–652.

[3] Milojevic-Dupont N, Wagner F, Nachtigall F, et al. EUBUCCO v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings. *Sci Data* 2023; 10: 147.

[4] Benjamin JR, Cornell CA. *Probability, Statistics, and Decision for Civil Engineers*. New York: Dover Publications Inc., 2014.

[5] Heeren N, Hellweg S. Tracking Construction Material over Space and Time: Prospective and Geo-referenced Modeling of Building Stocks and Construction Material Flows. *Journal of Industrial Ecology* 2019; 23: 253–267.

[6] Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California: Morgan Kaufmann Publishers Inc., 1988.

[7] Vatteri AP, D'Ayala D, Gehl P. Bayesian networks for assessment of disruption to school systems under combined hazards. *International Journal of Disaster Risk Reduction* 2022; 74: 102924.

[8] Mathern A, Steinholtz OS, Sjöberg A, et al. Multi-objective constrained Bayesian optimization for structural design. *Struct Multidisc Optim* 2021; 63: 689–701.

[9] Statistics Norway. 03158: Existing building stocks. All buildings, by region, type of building, year and contents. Statbank Norway. *SSB*, https://www.ssb.no/en/system/ (2023, accessed 27 April 2023).

[10] Geodata Norway. Eiendom | Geodata Online, https://dokumentasjon.geodataonline.no/docs/Temakart/Eiendom/ (2023, accessed 7 December 2023).

[11] Trondheim Kommune. Kulturminnekartet, https://www.trondheim.kommune.no/tema/bygg-kart-og-eiendom/byantikvar/aktsomhetskart-kulturminner (2023, accessed 6 November 2023).

[12] Statistics Norway. Standard for bygningstype / Matrikkelen, https://www.ssb.no/klass/klassifikasjoner/31# (2023, accessed 6 November 2023).

[13] European Environment Agency (EEA). *Modelling the Renovation of Buildings in Europe from a Circular Economy and Climate Perspective — European Environment Agency*. File, https://www.eea.europa.eu/publications/building-renovation-where-circular-economy/modelling-the-renovation-of-buildings/view (2022, accessed 1 December 2023).

[14] Ghione F, Mæland S, Meslem A, et al. Building Stock Classification Using Machine Learning: A Case Study for Oslo, Norway. *Front Earth Sci* 2022; 10: 886145.

[15] Gontia P, Nägeli C, Rosado L, et al. Material-intensity database of residential buildings: A case-study of Sweden in the international context. *Resources, Conservation and Recycling* 2018; 130: 228–239.

[16] Bergsdal H, Brattebø H, Bohne RA, et al. Dynamic material flow analysis for Norway's dwelling stock. *Building Research & Information* 2007; 35: 557–570.

[17] Lanau M, Liu G, Kral U, et al. Taking Stock of Built Environment Stock Studies: Progress and Prospects. *Environ Sci Technol* 2019; 53: 8499–8515.

[18] Ankan A. Exhaustive Search — pgmpy 0.1.23 documentation, https://pgmpy.org/structure_estimator/exhaustive.html (2023, accessed 5 December 2023).

[19] Ankan A, Panda A. PGMPY: Probabilistic graphical models using python. In: *Proceedings of the 14th python in science conference (scipy 2015)*. Citeseer, 2015.

[20] Statistics Norway. 03175: Existing building stocks. Residential buildings, by type of building (M) 2001 - 2023. Statbank Norway. *SSB*, https://www.ssb.no/en/system/ (2023, accessed 11 December 2023).

[21] Bache-Andreassen L. *Harvested wood products in the context of climate change*. 2009.