Probabilistic inference of material quantities and embodied carbon in building structures

Bernardino D'Amico[a,*], Jay H. Arehart[b]

[a] *School of Computing, Engineering & Built Environment – Edinburgh Napier University, UK*
[b] *Preoptima Ltd., Cambridge, UK*

Abstract

In an effort to minimise the carbon footprint of building structures, a range of prediction tools and methods have been recently proposed, so to enable design practitioners evaluating how their design choices ultimately affect the carbon embodied in their designs. Such tools are most often targeted for use at the early stage of the design process, that is when exploration of alternative design options is usually undertaken, hence room for potential carbon reductions is greatest and at no extra cost of redesign. The overarching methodology behind existing tools predominantly relies on idealised models to characterise the structural system, usually employing closed-form design equations and/or numerical Finite Element to generate an inventory of material quantity data (that is ultimately required for embodied carbon estimates). Despite the very high level of complexity achieved by some models, the absence of any empirical reference with 'as-built' inventory data of material quantities leaves room for doubt on how accurate such models really are in capturing the complexities and inherent variability of the population of real building structures such models aim to represent. To bypass this limitation, a data-driven probabilistic graphical model is proposed here as alternative to existing approaches. A Bayesian Network was developed and tested as a proof of concept, trained on a dataset of 133 data-points of real building structures, leveraging on six design variables (at most) to fully characterize the entire design space of early design options. Despite the very small set of 'explanatory' design variables, the model exhibited a 73% accuracy (mean average absolute prediction error of 27%) when predicting the embodied carbon on a test sample of unseen real building structures. The study ultimately demonstrates the viability of adopting a probabilistic (data-driven) approach for such an inference task as an inherently robust alternative to data-blind models currently proposed in literature.

Keywords: Embodied Carbon, Building Structures, Early Design Stage, Bayesian Network

1. Introduction

Building construction has been amply recognised as a major contributor to atmospheric green house gases (GHG) emissions [1] thus representing a major concern for it being a leading contributor to the climate crisis. Over the past decades, efforts from research, industry and

[*] Corresponding author
Email address: B.DAmico@napier.ac.uk (Bernardino D'Amico)

Preprint submitted to journal *May 20, 2024*

policymakers have primarily focused attention on reducing GHG emissions occurring during the operational stage of a building life-cycle. As a result, buildings have become increasingly more efficient to operate, thus leading in more recent times to a emphasis shift on the growing significance of embodied carbon impacts, which increasingly constitute a larger proportion of the overall life-cycle impacts attributable to buildings [2].

Among the various subsystems of a building, the structural system often accounts for the highest amount of embodied carbon impacts due to its substantial contribution to the overall building mass [3]. Improving design, with a focus on enhanced efficiency and more effective resource utilization, emerges as a promising strategy to mitigate embodied carbon in buildings. The importance of good design choices as early as possible during the design process has been amply acknowledged by design and research communities [4]. Consequently, a plethora of tools, methods, and strategies have been proposed for adoption by structural design practitioners.

While it is out of this study's scope to provide a comprehensive appraisal of all individual contributions reporting tools/findings on early design stage estimation of material quantities and/or associated carbon emissions in building structures —the reader is referred to recent work of Fang et al [5] for an up-to-date review— a high-level categorisation around two main general threads may be attempted in here, namely: *data-driven* investigations and *synthetic* numerical models.

With data-driven investigations we intend here the thread of studies aimed at characterising the embodied carbon of building structures by leveraging on inventory data of material quantities collected from case studies of real buildings [6, 7, 8, 9].

Although faithful to ground-truth information of 'as-built' material quantities, such approaches are inevitably based on a limited number of surveyed cases, therefore making hard to extrapolate their findings and conclusions beyond the specificity of the individual context being analysed. Notwithstanding the accumulation of ground-truth data and the qualitative insights they provide to a growing body of literature, the inherent limitedness of the analysed samples prevents a generalisation of the quantitative findings therein, and hence for them to be used as a prediction tool in other design contexts.

On a somehow opposite side, a perhaps more prolific thread of research studies is materialising. Primarily aimed at answering the same overarching question of how embodied carbon is later impacted by early design choices, such studies tackle the question by leveraging on a different approach to field data collection, that is, generating synthetic data of material quantities (and hence carbon) using a numerical model simulation of the structural design process. The underpinning implicit assumption justifying the sound validity of generating synthetic data to use for investigation is that any relevant feature of the built manufact can be traced back to a certain point along the design (and construction) process, and since building structures (unlike naturally occurring phenomena) are the result of a human process, this very process can be simulated with a reasonably high level of fidelity by informing the simulation model with knowledge of how building structures are designed in practice.

The range of research works in literature has grown over the last few years, both in complexity and scope. A wide range of studies employing synthetic numerical models of the structural design process to derive inventory of material quantities can be found. Studies may be focused to specific building types, e.g. to tall buildings [10, 11] or have scope narrowed to a sub-system of the whole structure, e.g. the super-structure [12], the floor slabs [13, 14, 15], single structural components such as beams [16, 17], or they only look at a certain material or construction technologies such as reinforced concrete [18], steel [19] or mass timber [20] or compare across all three such alternatives [21, 22]. Increasingly more complex models can be found in recent studies, expanding the modeling scope to both below- and above-ground parts of the structure, thus including foundations [23] and lateral load resisting system [10] as well as extending the model to other variables of interest other than embodied carbon, e.g. to operational carbon [24] and construction cost [23], and also going beyond the (frequent) modeling assumption of cuboid building shapes [23, 25].

The clear advantage of using a numerical model to generate synthetic material inventory data, is because of the possibility it entails to generate (theoretically unlimited) as many data as needed to characterise the statistical distributions of material quantities and related embodied carbon, hence enabling uncertainty quantification of results. In addition to bypassing the data-scarcity issue, a second advantage of relying on numerical models generating synthetic data is the possibility to tailor the model boundaries to a specific sub-domain of interest of the design process, or expanding the domain to account for as many variables of interest as possible —as evidenced by the above-mentioned research works— without having to worry about data availability and collection in the first place.

1.1. Why probabilities

While synthetic numerical modelling approaches provide a mean for assessing in a quantifiable way the influence of design variables (choices) on embodied carbon, the accuracy of their results is accepted insofar one is also accepting the model being an accurate representation of the real design (and construction) process being modelled. This is simply because, unlike statistical data-driven approaches, numerical models bear no connection with empirical data measurements of the variables they aim to predict, namely material quantities (and embodied carbon by proxy). Roughly put, to 'trust' the results one must first have 'faith' in the model assumptions.

A scientifically rigorous way to assess any kind of model is indeed by comparing its output with experimental results. Translated to our specific context, this would involve assessing the numerical model uncertainty (and hence its accuracy) by benchmarking the synthetically generated data against ground-truth field data such as 'as-built' material quantities —something done very seldom within the existing literature of early stage carbon prediction tools of building structures. What is often being assessed is instead the model sensitivity of results to the assumption of various model's parameters, that is: evaluating how the choice of assigning different values to (uncertain) variables in the model —either randomly drawn from a weighted distribution (often uniform) or a finite set of values— it is affecting the predicted material quantities and embodied carbon results, effectively moving from a deterministic model to a stochastic one.

Albeit some of the more sophisticated model frameworks described in the above-referenced studies adopt a stochastic approach to quantify the effect of parameters' uncertainty, with statistical distributions of materials and carbon being outputted instead of a single numerical value (as in a purely deterministic model), relations between variables remain essentially deterministic in nature, defined 'a priori' by the modeller, often using hard-coded sets of rules to describe how variables relate to each other in the model.

In a probabilistic model on the other hand, relations between variables are encoded by statistical frequency patterns learned directly from the collected data. Relations between variable are thus not defined in a deterministic fashion, tailored by the modeller, but probability distribution are used instead, learned from ground-truth data, therefore enabling the model to 'capture' both the interaction between variables in the real data-generating process as well as external influences from latent variables that are not explicitly modelled but 'present' in the collected data nonetheless.

To provide a simplified example of the conceptual and methodological differences between a probabilistic framework and the deterministic/stochastic approach often found in literature: the material intensity X of all structural members in a population of gravity frames is inevitably modelled as a function of some other variables, among which, the grid spans Y, and building type Z (a variable dictating the design floor loads). Drawing upon expert knowledge of the structural design process, such a functional relation $X = f(Y, Z)$ is modelled either using a set of closed-form mathematical equations or a numerical method (e.g. Finite Elements); in both cases the aim being to 'mimicking' the logical process of designing the structural frame as followed by practitioners in a real structural design setting. While such a functional relation may be pre-determined at large by the set of existing 'hard' rules a practitioner must follow by when it comes to design structural frames —i.e. as per requirements set out in design/construction standards and regulations in primis— there will always be other factors affecting the material intensity X which are not explicitly captured by the rule-based model. Empirical survey studies have for instance highlighted behavioural factors playing a major role on how building structures are designed in practice —e.g. the practitioners' tendency to over-engineering [26, 27, 28], or the clients' tendency to over-specify floor loads requirements [29]. On a more fundamental level, there is a practical impossibility to explicitly model the multidimensional space, inherent heterogeneity and context-specificity of the real and 'wider' population of building structures. As such, any rule-based numerical modeling approach demands simplifying assumptions to be made, which is not a negative per se, insofar the impact of such simplifications on the accuracy of the output prediction is somehow benchmarked against ground-truth data, so to assure the model is wrong, yet useful.[1] However this is very seldom the case in scientific literature of early stage carbon prediction tools of building structures.

A probabilistic framework enables to bypass such a methodological limitation in that relations between variables are learned directly from collected ground truth data. Returning on the gravity frame example: the relation between material intensity X as a function of grid span Y and building type Z is expressed in terms of probability distributions:

$$f(Y, Z) := P(X|Y, Z) \tag{1}$$

that is, expressing the probability of material intensity X given that we know the grid span $Y = y$ and building type $Z = z$ returns a distribution of probabilities for X instead of a single, uniquely determined value x.

Of course the likely reason why no attempts to train a probabilistic inference model from field data can be found in literature to date is because of the lack of a sufficiently large dataset

[1] "All models are wrong, but some are useful" is a famous aphorism coined by the British statistician George Box.

of material quantities of real building structures. Such a dataset was finally released by the UK structural design firm Price & Meyers and made publicly available on their website [30] as part of the their commitment to the Climate and Biodiversity Emergency Declaration.

Arguably, there exist a wide range of machine learning algorithms purposefully designed for prediction tasks, among which deep artificial neural networks (ANN) are standing out as perhaps the most important breakthrough in AI of recent years [31]. Yet, ANNs remain essentially opaque black-boxes, and hence they may be preferred as long as predictive accuracy is the main priority (e.g. over explainability). Conversely, probabilistic models maintain the same explicit representation of rule-based models on how variables relate to each other, as shown in Eq. (1) therefore making them easier to interpret.

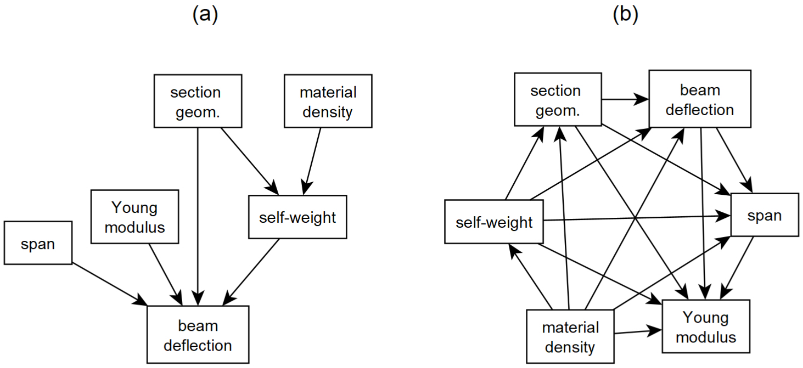

Figure 1: A simplified Bayesian Network example.

2. Proof of concept

To provide a tangible example of how the probabilistic approach can be successfully leveraged upon, the following sections describe the development and implementation of a Bayesian Network (BN) specifically designed as a decision support tool for inferring how (early) design choices influence the embodied carbon of building structures. Bayesian Networks are a particular instance of the more general AI class of probabilistic graphical models [32], enabling to represent knowledge of a given domain (along with its uncertainty) via a directed acyclic graph (DAG). Each node in the graph corresponds to a random variable of the system being modelled, whereas links connecting pairs of variables represent the probability distributions of the variable being connected (child) conditional to the set of connecting variables (parents). Such distributions can be stored in a tabular format when dealing with discrete variables. In this case, a BN is therefore fully defined by a set of conditional probability tables (CPTs) and a corresponding DAG.

Such a graphical representation, with nodes representing variables and links representing their probabilistic dependencies, it is a common feature of all probabilistic graphical models. BNs differentiate in that causal relationships between variable pairs are explicitly represented via directed links (e.g. unlike for Markov Random Fields) thus making them a more suitable option when modelling systems where these cause-effect relations are either entirely or partially understood, or can be elicited from domain experts.

BNs can be viewed as a ‘device’ encoding knowledge about a process or system in a compact representation form. To use an example the reader will be familiar with, the graph in Figure 1-a shows a simplified BN model of the deflection process of a structural beam, defined by a total of six X variables whose values x were generated by repeatedly performing a number of experimental tests on beams with different spans, Young modulus, section geometry and material density. In the absence of any ‘prior’ knowledge enabling us to rule out an existing dependency[2] between pairs of variables (i.e. as shown in Figure 1-b) the full joint probability distribution of the model would equate to the incremental product of each variable’s probability distribution conditional on every other variable (chain rule):

$$P(x_1, ..., x_6) = \prod_{i=1}^{n=6} P(x_i | x_1, ...x_{i-1}) \tag{2}$$

Conversely, expert knowledge tells us that *Young modulus* and *span* length are two random variables independent from each other, hence allowing us to rule out the existence of a link (statistical dependency) between the two —as well as between *section geometry* and *mat. density*, for that matter. This translates into a factorisation of the full joint where probability distributions of each variable are conditional on its parents only [33]:

$$P(x_1, ..., x_6) = \prod_{i=1}^{n=6} P(x_i | parents(X_i)) \tag{3}$$

which greatly reduces the number of instantiation entries (rows) in the conditional probability table of each variable. The advantage of having a sparser graph with fewer links is therefore that a more compact representation of the full joint probability distribution of all variables involved it is achieved, which is what enables compression of knowledge in Bayesian Networks —a key features of intelligent systems [34].

2.1. Data source and preprocessing

In order to learn the CPT for each variable in the Bayesian Network (also called model parameters), a training dataset of existing building structures is required. The publicly available dataset released by the Price & Meyers design firm (P&M) was employed for this proof of concept.

The latest dataset release [30] now in its 3rd edition, comprises more than 400 building structure data-points, however it was decided to use a previous yet smaller dataset version [35] as this latter is also reporting material quantities in addition to embodied carbon values. Out of the 275 individual building structure data-points contained in the older dataset version, only 78 were initially selected for training as they reported ‘as-built’ material quantities of new builds (hence excluding retrofit projects). This assured the model be trained on highly accurate material data measurements instead of design estimates.

However a preliminary inspection of this high quality data-subset soon revealed the number of as-built data-points being insufficient to cover for all the joint events expected to be

[2] In probability theory, two events (variables) are dependent when the outcome (value) of one event influences the outcome of the other event.

observed to fill the conditional probability tables. A joint event refers to the occurrence in the dataset of one or more variables taking on a certain assignment value, e.g. given a child variable X with parents Y and Z respectively taking assignment values x, y and z, the CPT's entry value $P(x|y, z)$ is obtained as a ratio between the observations' count of the event $\{x \wedge y \wedge z\}$ and the event's count of $\{y \wedge z\}$:

$$P(x|y,z) = \frac{P(x,y,z)}{P(y,z)} = \frac{count(x,y,z)}{count(y,z)} \tag{4}$$

The absence of event $\{y \wedge z\}$ in the training dataset would result in a zero-count value for $count(y, z)$ in Eq. (4) therefore yielding an undefined conditional probability for $P(x|y, z)$. Three mitigation measures were undertaken to prevent this:

- The initial high quality dataset of as-built data-points was further integrated with an additional 55 data-points reporting estimated material quantities, broken down according to the following RIBA Plan of Work stages [36]: 29 stage-5; 23 stage-4; 1 stage-3 and 2 stage-2 data-points, thus totalling up to 133 individual building project in the final training data-set. The resulting decrement in data quality was ameliorated by the bulk of additional data-points reporting material quantities estimated at RIBA stage-5, that is the construction stage, i.e. when only minor discrepancies are expected between estimated and as-built quantities of materials.

- A range reduction was applied to some variables. For categorical variables this implied collapsing two values together into a single value, such as for the *Basement* variable, having values {*Full-footprint; Partial-footprint; None*} in the original dataset mapped to {*True; False*}, thus reducing its range from 3 to 2. For continuous variables such as material quantities the value range (theoretically infinite) was reduced via discretisation. Particular attention was paid in calibrating the number of bins of each variable, and hence their width: while a smaller bin-width mitigates information loss due to discretisation, empty bins would occur below a certain threshold, thus resulting in unobserved events when computing CPT entries (Eq. (4)).

- Two dummy variables were added to the model to map the six-value range of *Superstructure type* into three-value ranges. Specifically, each of the six types of superstructure systems were classified according to their self-weight per unit of floor area into three ordinal values, namely: high, medium and low weight. Similarly, the six superstructure types were also allocated three categorical values indicating where (most of) concrete is located within its sub-components. A summary of how *Superstructure type* is mapped to the two dummy variables is provided in Table 1.

The full list of variables in the final training dataset is provided in Table 2 along with the corresponding value ranges and values.

2.2. Bayesian Network: model set-up

The DAG shown in Figure 2 (alongside Table A.1 in Appendix A.1) provides an overview of the probabilistic dependencies among variables encoded in the Bayesian Network. A parent

Superstructure unit-weight – 3 val.(s)	**Superstructure type** – 6 val.(s)	**Concrete elements** – 3 val.(s)
high	RC-frame	Frame&Floors
high	Masonry&Concrete	Floors
medium	Masonry&Timber	None
medium	Steel-frame&Precast/Composite	Floors
low	Steel-Frame&CLT	None
low	Timber-Frame(Glulam&CLT)	None

Table 1: Mapping values between the variable *Superstructure type* and variables *Superstructure unit-weight* and *Concrete elements*. This latter variable indicates where (most of) concrete is located within the superstructure. The variable *Superstructure unit-weight* is instead a qualitative (ordinal) measure of the superstructure's self-weight per unit of floor area.

variable in the DAG is connected with a directed link (arrow) to a child variable, thus implying a probabilistic influence between the two, such for instance between the variables *Basement* and *Concrete qty.* ($B \rightarrow C$). For these two, we have that a greater overall quantity of concrete is a more likely outcome when *Basement = true* instead of = *false*. This can be mathematically stated as inequality equation between the conditional probabilities of *C=high* given B:

$$P(C = high|B = true) > P(C = high|B = false) \tag{5}$$

An 'ancestor' instead is a third variable which also influences the child variable but it does indirectly so, mediated by the parent. For example, the variable *Superstructure unit-weight* (S^{**}) has an indirect influence on *Concrete qty.* mediated through the variable *Foundations type* ($S^{**} \rightarrow F \rightarrow C$). As for the $B \rightarrow C$ link, the explanation is rooted in domain's knowledge: a heavyweight superstructure, e.g. RC-frame, demands an increase of the foundations' bearing capacity, hence increasing the likelihood of a certain type of foundation (e.g. piles instead of pads) which adds up to the odds for *Concrete qty.=high*. Formally:

$$\begin{aligned} P(F = pile\text{-}caps|S^{**} = high) > P(F = pile\text{-}caps|S^{**} = low); \\ P(C = high|F = piles) > P(C = high|F = pads) \end{aligned} \tag{6}$$

Following the same line of reasoning for the opposite outcome, *Concrete=low* becomes more likely for the event $S^{**} = low$. The aim in here is to infer the conditional probability distribution of *Concrete qty.*, and other materials, given a set of evidence variables such as *Superstructure type*. A detailed description is provided in subsection 2.3.

2.2.1. Selecting variables

A general question in designing Bayesian Networks is whether richer models are always to be preferred to models with fewer variables. Inclusion of any new variable requires first to establish the variable 'relevance' to the model, that is, establishing whether the variable 'interacts' with others in the model and the statistical dependencies involved. For Bayesian Networks these interactions are explicitly represented though the DAG. To this end, some of the variables reported in the original P&M dataset were preliminary discharged as they were either deemed not causally relevant, or they were found to bear no association (lack of correlation) with material quantities and/or with all other variables in the graph.

Variable	**Symbol**	**Type**	**Unit**	**Variable's range & values**
Masonry & Blockworks qty.	M	Ordinal	m^2/m^2	5 values (bins): (0-1.64); (1.64-3.29);...(6.59-8.24)
Steel (sections) qty.	S	Ordinal	kg/m^2	12 values (bins): (0-13); (13-26);...(143-156)
Timber (products) qty.	T	Ordinal	kg/m^2	4 values (bins): (0-67); (67-134);...(201-268)
Reinforcement qty.	R	Ordinal	kg/m^2	7 values (bins): (0-30); (30-60); ... (180-210)
Concrete qty.	C	Ordinal	kg/m^2	18 values: (320-451); (451-582); ... (2540-2604)

GIFA	G	Ordinal	m^2	16 values (bins):
				(700-2000);
				(2000-3300);
				...
				(20200-21500)

Foundations type	F	Categorical	–	3 values:
				Piled(Ground-Beams/Caps);
				Mass(Pads/Strips);
				Reinforced(Pads/Strips/Raft)

Concrete elements	C^*	Categorical	–	3 values:
				Frame&Floors;
				Floors;
				None

Basement	B	Categorical	–	2 values: True; False

Superstructure type	S^*	Categorical	–	6 values:
				RC-frame;
				Steel-frame&CLT;
				Steel-frame&Precast/Composite;
				Timber-frame(Glulam&CLT);
				Masonry&Timber;
				Masonry&Concrete

Superstructure unit-weight	S^{**}	Ordinal	–	3 values:
				low;
				medium;
				high

No. storeys	N	Categorical	–	3 values:
				1-3 storeys;
				4-6 storeys;
				7-10 storeys

Cladding type	C^{**}	Categorical	–	2 values: Masonry; Other

Table 2: List of variables with corresponding ranges and values used to build the Bayesian Network for querying probability distributions of material quantities. Note: these latter are normalised per unit of gross internal floor area.

The attentive reader looking at Figure 2 will have noticed that other relevant variables influencing the probability distribution of material quantities are at play in addition to those considered here. The grid-span of the gravity frame for instance it's known to affect material intensity of the superstructure [23, 22]. Soil type, and its bearing capacity, are also some others *latent* variables, known to influence the type of foundation design which in turn affects the amount of concrete required overall.

While adding more 'relevant' variables generally improves the model fitting of the data —that is, its ability to make accurate and precise inference— as it is often the case, data required to represent such variables are either unavailable (as in this case) or unobservable.

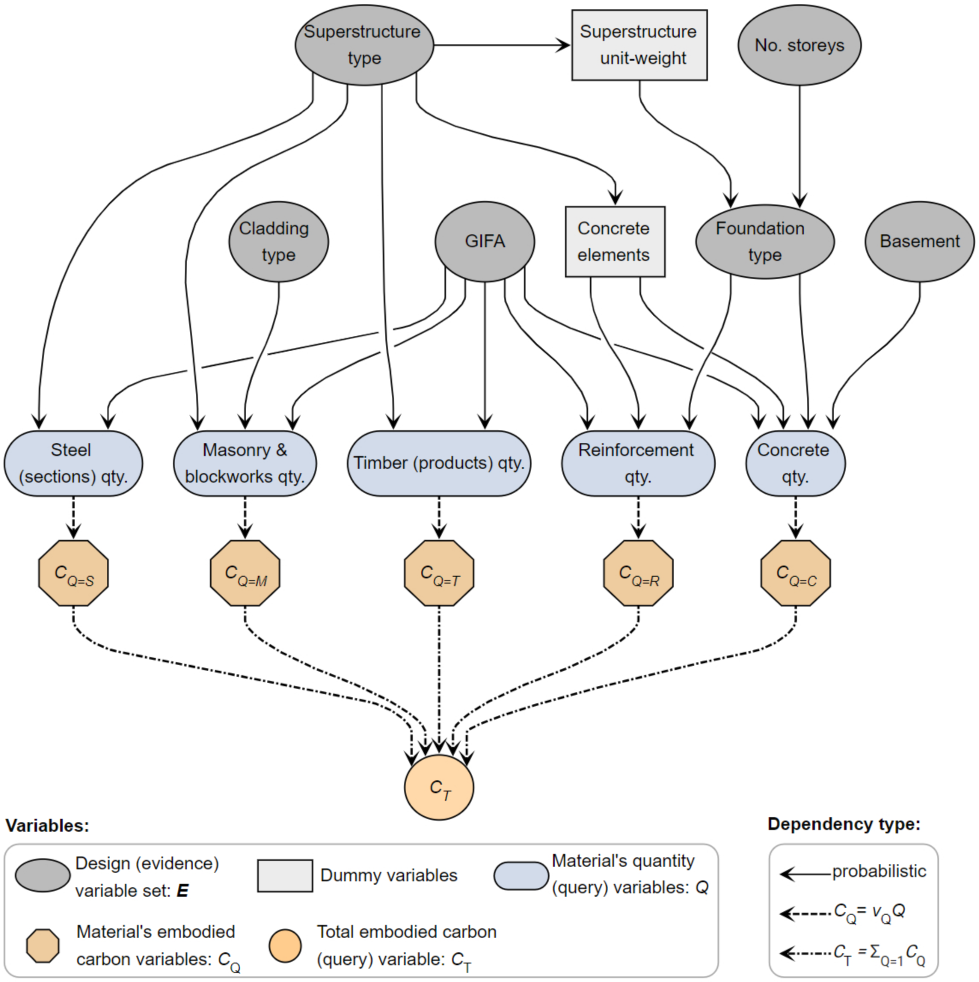

Figure 2: Directed acyclic graph (DAG) of the Bayesian Network employed for the proof of concept: Design information is encoded in the Network as *evidence* variables, whereas material quantities (normalised per unit of gross floor area) and total embodied carbon are *query* variables to infer.

According to Koller and Friedman however, it is not necessary to include every variable that might be relevant [33] insofar positive as well as negative influences affecting the variable's probabilities are accounted for. Let consider for example the variable *Foundations type* one more time: although the type of foundations may be negatively affected by whether the soil has a low bearing capacity, the probabilities already account for the fact that piled foundations may be needed despite a lightweight timber superstructure is bearing onto them (e.g. due to a high number of storeys, N):

$$P(F = piles|N = 10) > P(F = piles|N = 3) \tag{7}$$

This can be visualised in Figure 3, which shows the conditional probabilities found in the P&M dataset for *Foundations type* given *No. storeys* and *Superstructure unit-weight*, thus supporting the belief assumptions stated in eq. (7) and the first of eqs. (6) about the probability of piled foundations increasing with the number of storeys as well as with the type of gravity frame superstructure —e.g. there is less than 30% chance of piled foundations being used in conjunction with a low unit-weight superstructure such as timber-frame. This goes up to approx. 75% when a RC-frame is chosen instead.

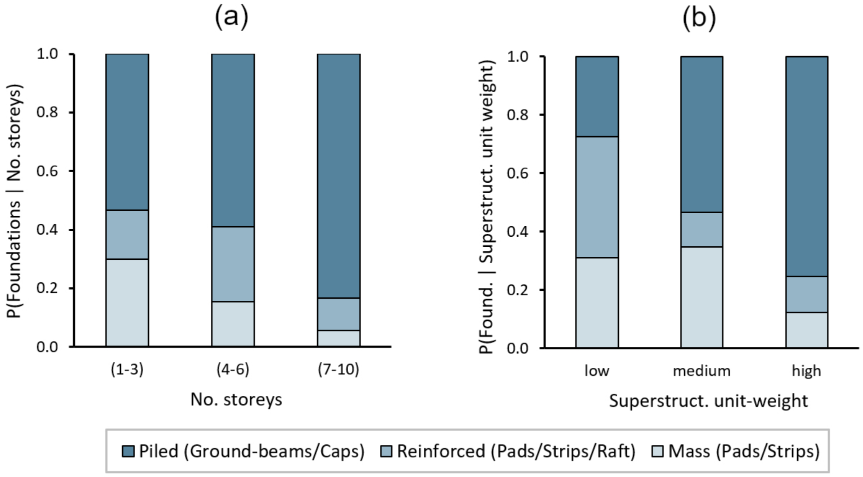

Figure 3: (a) Conditional probability for the variable *Foundations type* given *No.Storeys* and (b) *Foundations type* given *Superstructure unit-weight*. A trend in the data can be observed: the likelihood of piled foundations increases with the No. of storeys as well as with the unit weight of the superstructure.

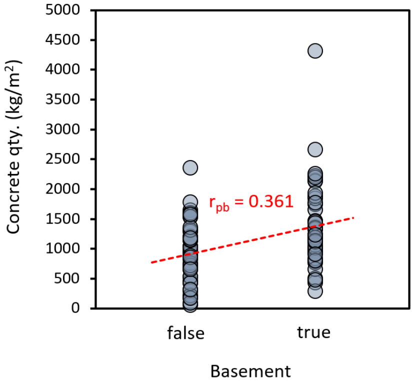

Figure 4: Point-biserial correlation (r_{pb}) between the variables *Basement* and *Concrete qty.*. The correlation trend is in line with existing findings [37].

2.2.2. Graph structure

Probabilistic interactions among variables were drawn from expert domain knowledge for this proof of concept, and thus graphically depicted in the graph in Figure 2 with the use of links with pointed arrows. The arrows' verse indicates the elicited causal effect between any two connected variables, e.g. for the variable pair $S^* - T$, it is the designer's choice of *Superstructure type* = *Timber-frame(Glulam&CLT)* that would 'cause' *Timber (products) qty.* to increase: $S^* \rightarrow T$, and not vice-versa.

In principle there is no practical or theoretical reason preventing to connect variables in the graph without following considerations of causal nature, as long as correlations between variables are captured. The supposed causal interaction *Basement* → *Concrete qty.*

underpinning eq. (5) implies that some statistical association (e.g. correlation) must also exist between the two variables —as it turns out to be the case by looking at the respective data plot (shown in Figure 4)— hence hinting at the possibility to learn the graph structure directly from the data [38]. Given however that the full space of existing graph structures grows super-exponentially with the number of variables [39], a brute-force approach of enumerating all DAGs to find the one that best fits the available data it is only feasible for a very limited number of variables. In practice, when dealing with large numbers of variables with no prior casual knowledge of the data-generating process, two main groups of algorithms are usually employed to learn the graph structure directly from data [40]: *score-based* or *constraint-based* algorithms (as well as hybridisations of the two). In score-based methods, finding the graph structure is essentially treated as an optimisation problem, that is to find the DAG (or set of DAGs) for which a scoring metric (set in terms of goodness of fit of the data) is being maximised. In constraint-based methods instead, the aim is to identify independencies among variables from the data using appropriate statistical independence tests (e.g. Chi-square for categorical variables) so that only graph structures encoding such independencies are searched for. Given for example two random variables X and Y, a test for absolute independence involves checking whether the joint distribution $P(X, Y)$ is significantly different[3] from the distribution one would expect were X and Y indeed absolutely independent:

$$P(X, Y) \approxeq P(X)P(Y) \Rightarrow X \perp\!\!\!\perp Y \tag{8}$$

or to test conditional independence between X and Y given a third variable Z:

$$P(X|Z) \approxeq P(X|Z, Y) \Rightarrow X \perp\!\!\!\perp Y|Z \tag{9}$$

To this end, the main advantage of structuring the DAG on considerations of cause-effect based on prior domain knowledge is that a sparser graph is obtained as opposed to linking together any pair of variables showing some correlation strength [33] thus achieving a more compact representation of the full joint probability distribution.

The reason why correlation may appear between two variables that are not sharing any obvious link of causality grounded in expert knowledge it is usually because of a third confounding variable causally interacting with the other two variables [41]. Here for instance a moderate negative correlation is observed in the data between the variables *Steel(sections) qty.* and *Reinforcement qty.* as shown in Figure 5-a, yet no plausible expert explanation of causality can be thought of between the two. However by isolating the following trail of interactions from the final graph:

$$Reinf.\ qty. \leftarrow Concrete\ elem. \leftarrow Superstr.\ type \rightarrow Steel(sections)\ qty. \tag{10}$$

it can be seen how the variable *Superstructure type* is a common cause to both *Steel (sections) qty.* and to *Reinforcement qty.* By controlling the data for the common-cause variable, the negative correlation disappears[4] as shown in Figure 5-b, with steel sections

[3] Based on a statistical *significance level* threshold value. The lower the threshold, the stricter the independence test.

[4] The phenomenon is known in statistics as the “Simpson’s paradox” [42].

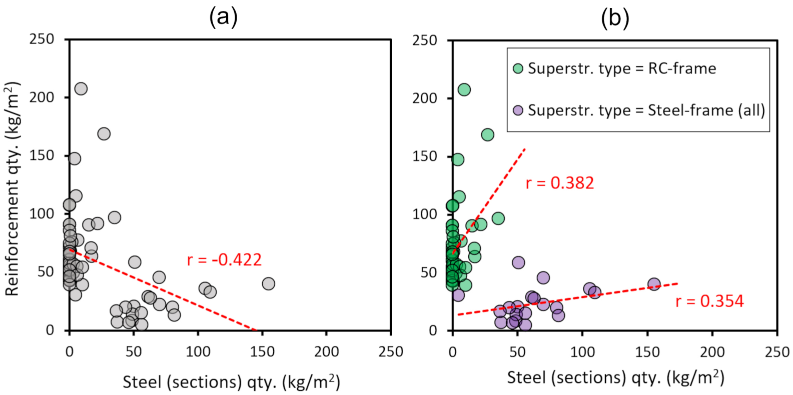

Figure 5: (a) A moderate negative Pearson's correlation ($r = -0.422$) is observed between two seemingly unrelated material quantity variables: *Steel (sections) qty.* and *Reinforcement qty.*; (b) The correlation is inverted after controlling for the confounding variable *Superstructure type*.

and reinforcement material intensities becoming positively correlated after controlling for *Superstructure type*. This was deemed enough indication to avoid placing a direct link in the graph between *Steel (sections) qty.* and *Reinforcement qty.*

2.3. Making probabilistic queries

The described BN set-up with trained parameters can finally be used to make probabilistic queries, that is, computing the posterior belief probability of a *query* variable Q (or more than one) taking on certain value q, or computing the probability distribution across the entire range of values of each Q. A query will usually involve some other variables in the model with observed assignments (so called *evidence* variables). With reference to Table 2 and Figure 2, query variables are the five material quantities: $Q \in \{S, M, T, R, C\}$ as well as the total embodied carbon variable C_T, whereas the set of evidence variables will be a subset of the six design variables: $\mathbf{E} \subseteq \{S^*, C^{**}, N, F, B, G\}$.

2.3.1. Inferring material quantities

The posterior probability distribution of each material variable Q conditional on a set $\boldsymbol{E}$ of design evidences, can be derived as a ratio between a joint and a marginal distribution:

$$P(Q|\boldsymbol{E}) = \frac{P(Q, \boldsymbol{E})}{P(\boldsymbol{E})} \tag{11}$$

Considering for example the query $P(c|s^*, n)$ to infer the probability of *Concrete qty.* $= c$ given *Superstructure type* $= s^*$ and *No. storeys* $= n$ as evidence, then the problem would be reduced to find the joint $P(c, s^*, n)$ and marginal $P(s^*, n)$:

$$P(c|s^*, n) = \frac{P(c, s^*, n)}{P(s^*, n)} \tag{12}$$

both of which could, in theory, be derived from the full joint distribution of all variables in the model by 'summing out' (marginalising) all the other variables that are not queries nor evidence.

Such an approach would clearly requires the full joint probability table to be explicitly available. Enumeration of all the entries of the full joint table could in theory be performed via Eq (3). In reality, expanding the full joint to answer queries it is either computationally inefficient or simply unfeasible due to the exponential blow up of entries in the full joint table as the number of variables increases. For a small sized model like the one in here —comprising of 13 variables with a total of 84 value states (excluding embodied carbon variables)— a table with 940'584'960 entry rows would be required to explicitly represent the full joint distribution.

A more computationally efficient approach that reduces both the number of entry values to store, as well as the required arithmetical operations, it is the Variable Elimination algorithm, which was adopted here to perform inference queries for all material quantity variables. Implementation details of how to compute the example query in Eq (12) via Variable Elimination algorithm can be found in Appendix A.2.

Both approaches of brute-force enumeration or using the Variable Elimination algorithm, are so-called *exact* inference methods in that probabilities are the direct result of a finite set of arithmetical operations. In situations where exact inference becomes computationally intractable, e.g. due to the model size and/or the graph complexity, approximate methods are employed instead to reconstruct the queried distribution [33], thus obtaining an approximated estimation of the probability estimate that gradually improves as sampling proceeds. A Montecarlo-based approximate method was used here to infer total embodied carbon from the material quantity distributions obtained via Variable Elimination, as described in the following subsection.

2.3.2. Inferring embodied carbon

For each material quantity variable Q in the P&M dataset used here to build the training dataset, a corresponding embodied carbon variable C_Q covering life cycle stages A1-A5 [43] was also provided (see DAG in Figure 2). As common in industry practice, the values c_Q of these embodied carbon variables C_Q were obtained by multiplying each material quantity $Q = q$ with a relevant carbon coefficient v_Q, that is, a multiplier estimated using life cycle assessment methodology (LCA) thus expressing the global warming potential in $kg_{CO_{2e}}$ per declared unit of material — e.g. unit of mass for *Steel (section) qty.* or unit of wall area for *Masonry & Blockworks qty.*:

$$\begin{aligned} c_Q[kg_{CO_{2e}}] = q[kg_{mat.}] \cdot v_Q[kg_{CO_{2e}}/kg_{mat.}]; \\ c_Q[kg_{CO_{2e}}] = q[m^2_{mat.}] \cdot v_Q[kg_{CO_{2e}}/m^2_{mat.}] \end{aligned} \tag{13}$$

Established the existence of this linear proportionality relation, $c_Q \propto q$, there is no intrinsic uncertainty as such about the embodied carbon of each material reported in the dataset (c_Q) since their probability distribution is effectively matching the distribution of the corresponding material quantity:

$$P(C_Q = c_Q|Q = q) = P(Q = q|\mathbf{E}) \tag{14}$$

where the assignment values c_Q in Eq.(14) matching each probability entry table of $P(C_Q|Q)$ can be derived straightforwardly via Eq. (13). According to the authors of the P&M dataset, the original carbon coefficients v_Q used to build the dataset were drawn from the ICE database [44] however the actual values for each material for each data-point were not explicitly reported, therefore they were back-calculated here for the purpose of this proof of concept by performing a linear regression of all the $\{c_Q, q\}$ value pairs in the original dataset (see Appendix A.3). Arguably, a more holistic approach to account for carbon intensities embodied in each material would be through a dedicated probabilistic model [45]. This is especially true if considering the degree of uncertainty surrounding carbon coefficients of construction materials [46, 47]. While superior, such an approach would entail integrating the proposed (probabilistic) model to infer material quantities with a probabilistic assessment models of each material's life cycle, thus requiring LCA domain expertise and a data collection of materials' processes and flows for model training: two kind of resources not always available to structural design firms. On the other hand, structural designers can more comfortably rely for their embodied carbon assessments on an ever-increasing number of county/region-specific databases of pre-compiled single-valued carbon multipliers and on a proliferation of product-specific EPDs[5] issued by materials' manufacturers —a single-valued approach this that is also endorsed by professional bodies [49]. Notwithstanding the higher prediction accuracy that could be achieved by accounting for the variability of carbon coefficients, the single-valued approach adopted here is ultimately instrumental to enable a comparison between predicted embodied carbon and the corresponding 'true' value in the P&M dataset (section 3).

Having obtained (a probability distribution of) the embodied carbon contribution of each material variable via Eq (14), the probability distribution of total embodied carbon, C_T, conditional on these variables is:

$$P(C_T|C_1, ..., C_Q, ..., C_5) \text{ where } Q \in \{S, M, T, R, C\} \tag{15}$$

and it can be computed via Montecarlo sampling. Specifically, five population samples of c_Q values are randomly drawn (one for each variable C_Q) using as probability weighting the previously found distributions $P(C_Q|Q)$. Values of the five populations so obtained are then combined together into a single population $\boldsymbol{c}$ of total embodied carbon values c_T:

$$c_T = \sum_{Q=1}^{n=5} c_Q \tag{16}$$

The discrete probability distribution (histogram) of total embodied carbon C_T in Eq. (15) is then constructed by inspecting the frequency of occurrence of c_Q values in the population dataset $\boldsymbol{c}$ generated via Eq. (16):

$$P(C_T = c_T) = \frac{count(c_T)}{|\boldsymbol{c}|} \tag{17}$$

As such, the end-to-end process of querying the total embodied carbon probability distribution $P(C_T|\mathbf{E})$ given one, more than one, or no design variables as evidence, it is carried out by first inferring distributions of material quantities $P(Q|\mathbf{E})$ via Variable Elimination and then applying Eqs. (13) to (17).

[5] Environmental Product Declarations [48].

3. Numerical results

The described probabilistic framework for material (quantities) intensities and resulting embodied carbon of building structures has been implemented into a computer application using Python programming language and it was used to generate all the numerical results described in this section. Implementation of the Variable Elimination algorithm, required for exact inference of material quantities, relies upon the `pgmpy` library by Ankan and Textor [50]. To enable repeatability of results, a repository containing all datasets and the Python source codebase is made available online [51].

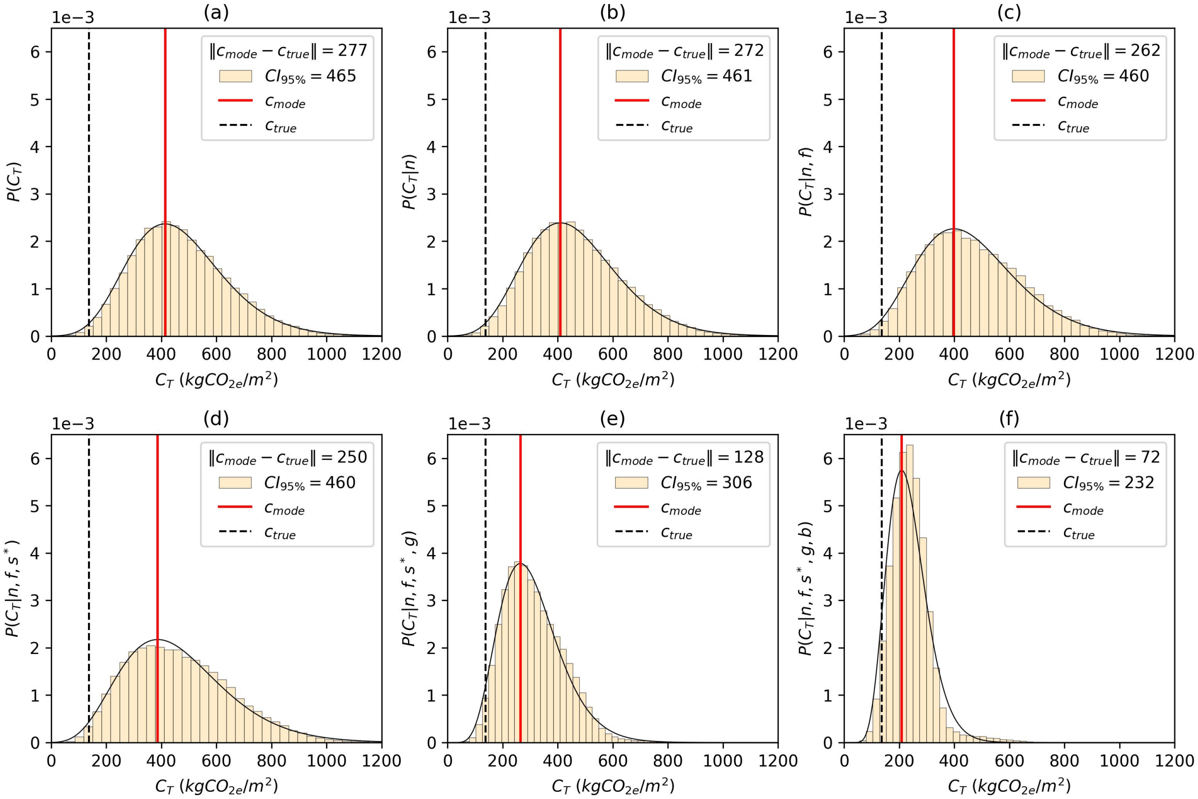

Figure 6: The inferred probability distribution of total embodied carbon C_T becomes more accurate and precise as evidence updates. The gap in accuracy between predicted (most likely) carbon value, c_{mode}, and true value, c_{true}, is narrowing as more design information is cumulatively provided for each query update as follows: (a) No design information; (b) No. storeys = between 1 and 3; (c) Foundations type = Mass(Pads/Strips); (d) Superstructure type = Timber-Frame(Glulam&CLT); (e) Building size (GIFA) = 3390 m^2; (f) Basement = false.

3.1. Accuracy and precision

In order to visualise how prediction accuracy of the probabilistic model improves with updating evidence, Figure 6 shows probability distributions (gamma-fitted histograms) of total embodied carbon, $P(C_T|\mathbf{E})$, obtained by running a series of queries with an increasingly larger set $\mathbf{E}$ of design evidence variables, starting with no evidence at all: $\mathbf{E} = \{\}$. Every time a design variable is added to the evidence set, its assignment value is matching the design value of a reference individual data-point picked from the training dataset. A reference design data-point with a low embodied carbon value, $C_T = c_{true} = 136.4\ kgCO_{2e}/m^2$ (dashed line in Figure 6) was chosen as it lies far from the mode (most likely value: c_{mode}) of carbon distribution initially inferred with no evidence at all: $P(C_T)$. A low probability for c_{true} is inferred by this initial query (Figure 6-a), yet as the carbon probability updates with more evidence (design) variables, the distance between *true* carbon value and inferred mode value narrows down, i.e. prediction accuracy improves. Of note, the accuracy distance $\|c_{mode} - c_{true}\|$ may not reduce incrementally with each new piece of evidence because (all other things being equal) a carbon increase may actually be a more likely outcome depending on the evidence value being assigned —such a sensitivity of prediction accuracy to the evidence ordering can be visualised in Figure 7. Nonetheless a clear trend in improved prediction accuracy can be observed: with reference to Figure 6, the initial deviation between true value and inferred value without prior evidence $\|c_{mode} - c_{true}\| = 277\ kgCO_{2e}/m^2$ narrows down to 72 $kgCO_{2e}/m^2$ when the belief probability of carbon is fully updated with values of five design variables. The probability distribution given *Cladding* is not shown in Figure 6 for the sake of formatting the Figure with six subplots instead of seven.

It is worth remarking that the 'true' embodied carbon value of the reference design data-point may not be an accurate representation of the real built project in that it was derived using single-valued carbon coefficients (see section 2.3.2) which provide an average estimate of carbon embodied in declared units of material. Nonetheless, as we aimed here at evidencing how prediction accuracy is improving with additional knowledge (evidence) of building design variables being fed to the probabilistic model, the closeness of prediction to the 'true' carbon value would improve regardless of the carbon coefficients being used insofar such design variables and materials carbon coefficients are independent (e.g. the number of storeys bears no influence on the energy mix supplied to manufacture steel sections, or vice-versa).

In addition to accuracy, precision also improves with updating evidence. With reference to Figure 6, the 95% confidence interval around the mean of the probability distribution ($CI_{95\%}$) is around 465 $kgCO2e/m^2$ for $P(C_T)$ i.e. when no design information is given. This reduces by half to 232 $kgCO2e/m^2$ for $P(C_T|N, F, S^*, G, B)$ i.e. when design information on five design variables is provided. A graphical overview of the probability update of material quantities can be found in Appendix A.4.

3.2. Model testing

In order to asses the model's extent to generalise on new data, the inference exercise previously described in subsection 3.1 was performed on a sample dataset of unseen building design data-points. This testing data-set was assembled by randomly selecting six design data-points from the preprocessed P&M dataset. All six design variables N, F, S^*, G, B and C^{**} were considered this time for evidence (i.e including *Cladding type*) hence yielding to $2^6 = 64$ individual queries per single data-point, that is a total of 384 inference queries. To get a quantitative indication of how prediction accuracy improves with number of evidence variables, the resulting absolute percentage errors ($\|c_{mode} - c_{true}\|/c_{true}$) were clustered together based on the number of evidence variables being provided for querying, hence going from no evidence: $|\mathbf{E}| = 0$, to the full evidence set: $|\mathbf{E}| = 6$. The mean absolute percentage error (MAPE) of each cluster is shown in Figure 8-a. Here it can be seen how the same trend in accuracy improvement —that was previously observed on a single data-point picked form the training dataset— it is also holding on a sample of unseen data. Specifically, the mean

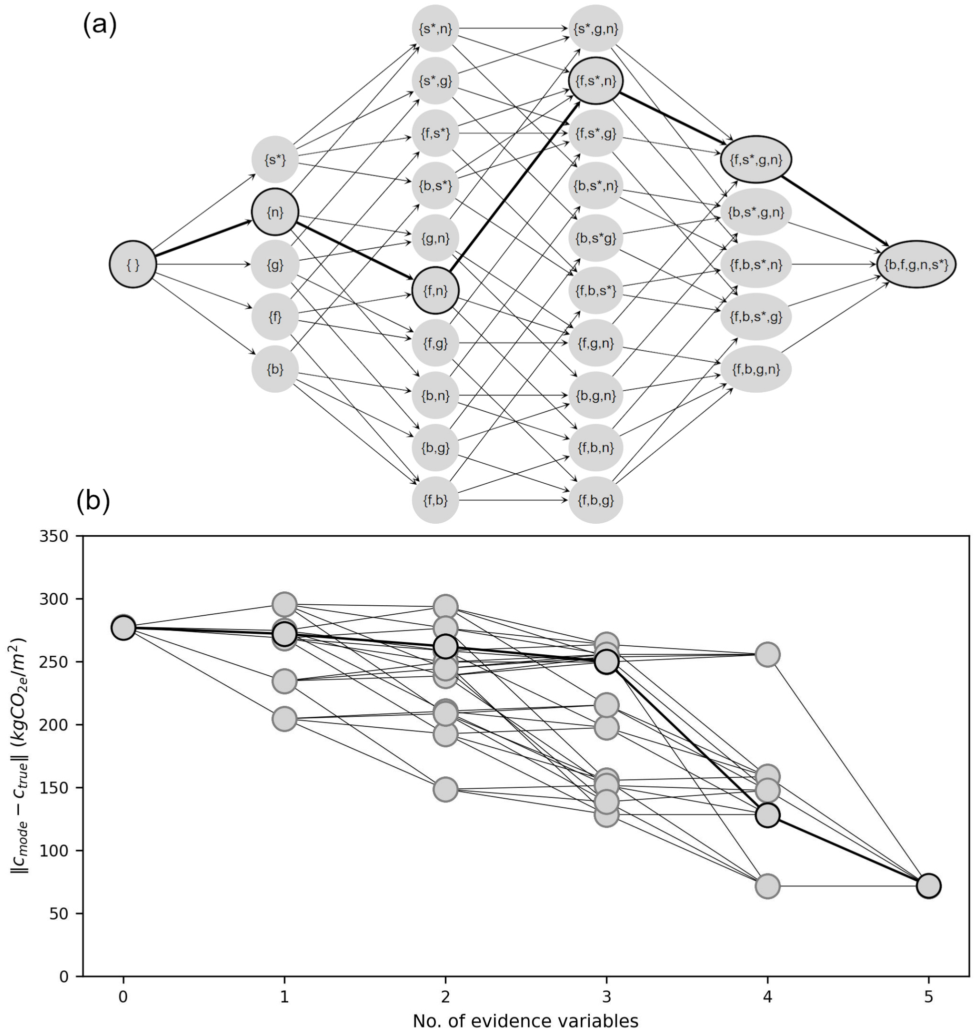

Figure 7: Sensitivity of prediction accuracy to the ordering of evidence (design) variables available for querying: (a) *Hasse* diagram [52] showing all possible evidence variables' orderings, starting with no evidence on the far left, $\mathbf{E} = \{\}$, to 'full' evidence on the far right, $\mathbf{E} = \{N, F, S^*, G, B\}$; (b) Estimation error $\|c_{mode} - c_{true}\|$ for each evidence variables set, represented as a node in the Hasse diagram. The bolt line indicates the ordering of evidence variable sets, $\mathbf{E}$, followed to generate the query results shown in Figure 6.

absolute prediction error of total embodied carbon reduces from about 43 % when no design evidence is given, down to circa 27 % given information for all six design variables. Notably, the spread of prediction error also tends to narrow down with increasing design information, as shown in Figure 8-b, where the amount of variation of the prediction error is reported in terms of its standard deviation.

3.3. Remarks and discussion

Whilst a prediction accuracy of 73 % might seem unimpressive at a first look, it is rather remarkable that it has been achieved by using as few as six variables to characterise the entire space of design features, covering for the whole building structure, both above- and below-ground. In the best knowledge of the authors, this is the first study of an early stage design tool with such a wide prediction scope that is also benchmarked for accuracy

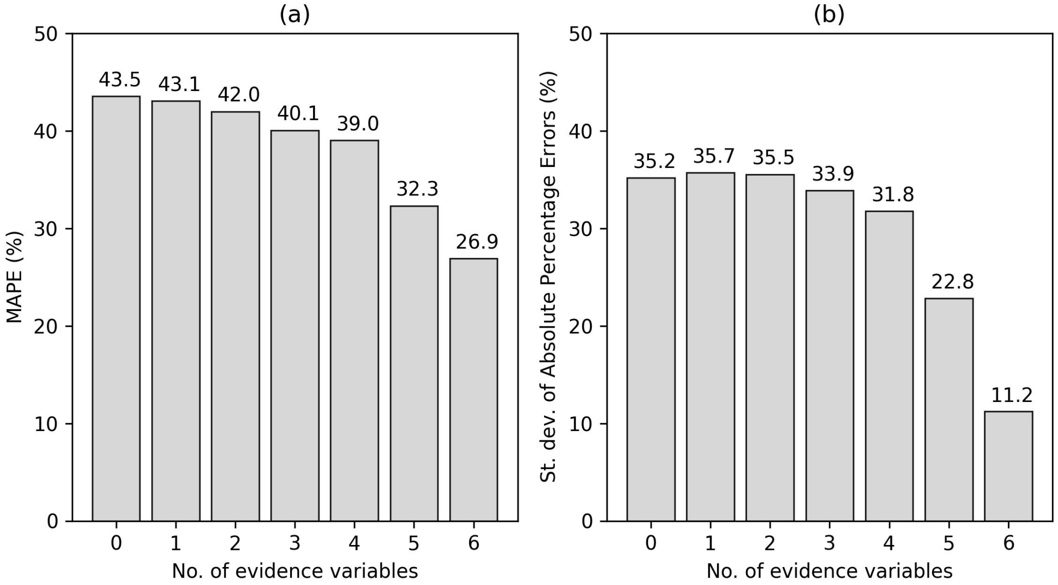

Figure 8: (a) Reduction trend of the mean absolute percentage error (MAPE) of predicted total embodied carbon for a sample population of unseen building design data-points; (b) Corresponding standard deviations of the absolute percentage errors.

against as-built material quantities. It is also worth to remind, the main aim of this proof of concept has been primarily set to showcase the untapped potential of employing probabilistic graphical models as an effective early design tool for inferring material quantities (and hence embodied carbon) of whole-building structures. Existing research works reporting findings on (or describing tools for) characterisation of material/carbon intensities are 'static' in nature, whereas the level of accuracy of a probabilistic data-driven model is scalable, hence only limited by the volume of data-points it is being trained on and number of relevant design variables it is accounting for.

Furthermore the described Bayesian Network framework can be straightforwardly applied to solve the inverse problem of inferring design parameters given a target carbon intensity, e.g. inferring the mostly likely number of storeys or superstructure type, or any other variable of design, setting a certain embodied carbon target value for the evidence variable (so called *diagnostic* modality in the relevant AI literature). In other words, by simply swapping the sets of query and evidence variables would suffice for using the exact same BN architecture as a generative design tool instead of an inference tool.

Lastly, as suggested by one of the reviewers, the described predictive framework could be used to develop predictive benchmarks of building material quantities, to be used for setting material efficiency (and whole life carbon) policy thresholds, albeit a coordinated effort would be preliminary required in order to collect and curate training datasets that are representative of the whole (e.g. county-wise) building design population.

4. Study limitations

Having showcased the untapped potential of probabilistic models to the specific prediction task of inferring material quantities and related embodied carbon in building structures, the main practical barrier for larger and more powerful models remains of course the availability of ground truth data required to train the model parameters. Despite the existing calls for more collaborative efforts to build shared repositories [53, 54], design firms are historically reluctant in open sourcing their in-house databases of past construction projects (with some rare and laudable exceptions such as Price & Meyers). Climate change is however a global crisis affecting us all, which can only be tackled effectively with concentrated joint efforts from all stakeholders involved.

5. Conclusions

Due to the major impact of building construction on climate change, recent research efforts have been focused to identify strategies and tools to inform and help design practitioners understanding (and hence mitigating) how their early design choices impact the embodied carbon of their service product, namely, building structures.

Such an effort has resulted in most (if not all) approaches, drawn from existing relevant literature on the subject, primarily focused on modeling the structural design process in a deterministic fashion, often leveraging on closed-form equations or numerical (Finite Element informed) models to generate inventories of material quantity data from which a metric of embodied carbon can then be estimated. While advantageous on many aspects, such methodological approaches relying on synthetically generated sets of data are rarely benchmarked for accuracy of prediction against ground truth field data of 'as-built' material quantities, hence leaving potential room for doubt regarding their ability to properly capture the complexities involved in the design and construction process of building structures, which is essential in order to provide design practitioners with a tool able to accurately predict the carbon eventually embodied in the built manufact.

Thanks to the release in 2022 of a relatively large dataset of real structural building projects by the UK firm Price & Meyers, it was possible here to investigate the feasibility of employing a probabilistic graphical model trained directly on collected material quantity data (mostly of which were from 'as built' measurements), therefore enabling to infer how early design choices influence the embodied carbon of building structures accounting in a natively robust and automated way for the inherent variability and heterogeneity of the target population being modelled.

This was achieved by building, running and testing a proof-of-concept Bayesian Network model employing a total of six 'explanatory' design variables to represent the full set of characteristic features usually considered at early design stage, namely: (I) The type of superstructure, (II) number of storeys, (III) type of cladding, (IV) the gross internal floor area (as a proxy of building size), (V) the type of foundation design and (VI) the presence/absence of a basement.

Despite the 'coarseness' of the model (relying on just six design variables, at most) when tested on a sample of unseen data-points, it was able to score a mean average absolute prediction error of 27 %, meaning that it was able to infer real quantities of all structural materials in the whole structure (both above- and below-ground) and carbon embodied therein with a prediction accuracy of 73 %, which is a rather remarkable result if considering that as fewer as six explanatory variables were used for evidence. Above all, introducing a ready-to-use tool for predicting material quantities and embodied carbon was a secondary objective of this study, with the primary aim being instead to showcasing the feasibility of approaching the problem with a probabilistic method of inference, so to bypass the inherent model fragility of deterministic rule-based methodologies.

6. Declaration of competing interest

The authors Bernardino D'Amico and Jay Arehart disclose that they are shareholders of shares in Preoptima Ltd., a company that has filed a patent related to the method described in this research paper.

7. Acknowledgements

The first author would like to acknowledge Prof. Peter Andras for providing insightful discussions and information sources on probabilistic graphical methods. Acknowledgement are also extended to Price & Meyers for their effort in assembling and publicly releasing the embodied carbon database of real building structures used here for training, and without which the presented work would not have been possible. We also thank the anonymous reviewers for their comments and feedback which greatly improved the final version of this paper.

8. Data availability

All data and code developed for this study is available on a GitHub repository: [51].

References

[1] L. F. Cabeza, Q. Bai, P. Bertoldi, J. Kihila, A. Lucena, E. Mata, S. Mirasgedis, A. Novikova, Y. Saheb, IPCC, Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (2022).

[2] M. Röck, M. R. M. Saade, M. Balouktsi, F. N. Rasmussen, H. Birgisdottir, R. Frischknecht, G. Habert, T. Lützkendorf, A. Passer, Embodied GHG emissions of buildings — The hidden challenge for effective climate change mitigation, Applied Energy 258 (2020) 114107.

[3] S. Kaethner, J. Burridge, et al., Embodied CO2 of structural frames, The structural engineer 90 (5) (2012) 33–40.

[4] C. F. Dunant, M. P. Drewniok, J. J. Orr, J. M. Allwood, Good early stage design decisions can halve embodied CO2 and lower structural frames' cost, in: Structures, Vol. 33, Elsevier, 2021, pp. 343–354.

[5] D. Fang, N. Brown, C. De Wolf, C. Mueller, Reducing embodied carbon in structural systems: A review of early-stage design strategies, Journal of Building Engineering (2023) 107054.

[6] W. Hawkins, S. Cooper, S. Allen, J. Roynon, T. Ibell, Embodied carbon assessment using a dynamic climate model: Case-study comparison of a concrete, steel and timber building structure, in: Structures, Vol. 33, Elsevier, 2021, pp. 90–98.

[7] R. Kumanayake, H. Luo, N. Paulusz, Assessment of material related embodied carbon of an office building in Sri Lanka, Energy and Buildings 166 (2018) 250–257.

[8] H. Li, Q. Deng, J. Zhang, B. Xia, M. Skitmore, Assessing the life cycle co2 emissions of reinforced concrete structures: Four cases from china, Journal of cleaner production 210 (2019) 1496–1506.

[9] J. Monahan, J. C. Powell, An embodied carbon and energy analysis of modern methods of construction in housing: A case study using a lifecycle assessment framework, Energy and buildings 43 (1) (2011) 179–188.

[10] J. Helal, A. Stephan, R. H. Crawford, The influence of structural design methods on the embodied greenhouse gas emissions of structural systems for tall buildings, in: Structures, Vol. 24, Elsevier, 2020, pp. 650–665.

[11] P. Foraboschi, M. Mercanzin, D. Trabucco, Sustainable structural design of tall buildings based on embodied energy, Energy and Buildings 68 (2014) 254–269.

[12] M. P. Drewniok, J. Campbell, J. Orr, The lightest beam method–a methodology to find ultimate steel savings and reduce embodied carbon in steel framed buildings 27 (2020) 687–701.

[13] J. M. Broyles, J. P. Gevaudan, M. W. Hopper, R. L. Solnosky, N. C. Brown, Equations for early-stage design embodied carbon estimation for concrete floors of varying loading and strength, Engineering Structures 301 (2024) 117369.

[14] A. Jayasinghe, J. Orr, T. Ibell, W. P. Boshoff, Minimising embodied carbon in reinforced concrete flat slabs through parametric design, Journal of Building Engineering 50 (2022) 104136.

[15] B. D'Amico, F. Pomponi, J. Hart, Global potential for material substitution in building construction: The case of cross laminated timber, Journal of Cleaner Production 279 (2021) 123487.

[16] M. Sahebi, M. Dehestani, Sustainability assessment of reinforced concrete beams under corrosion in life-span utilizing design optimization, Journal of Building Engineering 65 (2023) 105737.

[17] A. Jayasinghe, J. Orr, T. Ibell, W. P. Boshoff, Minimising embodied carbon in reinforced concrete beams, Engineering Structures 242 (2021) 112590.

[18] R. Keihani, M. Tohidi, A. Janbey, A. Bahadori-Jahromi, Enhancing sustainability of low to medium-rise reinforced concrete frame buildings in the uk, Engineering Future Sustainability 1 (1) (2023).

[19] B. D'Amico, F. Pomponi, Accuracy and reliability: A computational tool to minimise steel mass and carbon emissions at early-stage structural design, Energy and Buildings 168 (2018) 236–250.

[20] J. M. Greene, H. R. Hosanna, B. Willson, J. C. Quinn, Whole life embodied emissions and net-zero emissions potential for a mid-rise office building constructed with mass timber, Sustainable Materials and Technologies 35 (2023) e00528.

[21] J. Hart, B. D'Amico, F. Pomponi, Whole-life embodied carbon in multistory buildings: Steel, concrete and timber structures, Journal of Industrial Ecology 25 (2) (2021) 403–418.

[22] B. D'Amico, F. Pomponi, On mass quantities of gravity frames in building structures, Journal of Building Engineering 31 (2020) 101426.

[23] H. Gauch, W. Hawkins, T. Ibell, J. Allwood, C. Dunant, Carbon vs. cost option mapping: A tool for improving early-stage design decisions, Automation in Construction 136 (2022) 104178.

[24] H. Gauch, C. Dunant, W. Hawkins, A. C. Serrenho, What really matters in multi-storey building design? a simultaneous sensitivity study of embodied carbon, construction cost, and operational energy, Applied Energy 333 (2023) 120585.

[25] A. Stephan, F. Prideaux, R. H. Crawford, Epic grasshopper: A bottom-up parametric tool to quantify life cycle embodied environmental flows of buildings and infrastructure assets, Building and Environment (2023) 111077.

[26] M. C. Moynihan, J. M. Allwood, Utilization of structural steel in buildings, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 470 (2168) (2014) 20140170.

[27] A. Alnuaimi, M. Al Mohsin, A. Hago, S. El Gamal, Overdesign of villa structures in Oman, The Journal of Engineering Research [TJER] 12 (2) (2015) 68–76.

[28] C. F. Dunant, M. P. Drewniok, S. Eleftheriadis, J. M. Cullen, J. M. Allwood, Regularity and optimisation practice in steel structural frames in real design cases, Resources, Conservation and Recycling 134 (2018) 294–302.

[29] J. Orr, M. P. Drewniok, I. Walker, T. Ibell, A. Copping, S. Emmitt, Minimising energy in construction: Practitioners' views on material efficiency, Resources, Conservation and Recycling 140 (2019) 125–136.

[30] B. Gholam, Price & Myers Embodied Carbon Database - 3rd edition (2023).
URL https://www.pricemyers.com/wo_files/files/PM_Carbon_Database_for_Release_2023_V2.pdf

[31] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.

[32] F. V. Jensen, T. D. Nielsen, Bayesian networks and decision graphs, Vol. 2, Springer, 2007.

[33] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.

[34] P. Maguire, O. Mulhall, R. Maguire, J. Taylor, Compressionism: a theory of mind based on data compression, in: Proceedings of the 11th International Conference on Cognitive Science, 2015, pp. 294–299.

[35] B. Gholam, Price & Myers Embodied Carbon Database - 2rd edition (2022). URL https://www.pricemyers.com/wo_files/files/PM_Carbon_Database_11-02-2022.pdf

[36] Royal Institute of British Architects, RIBA Plan of Work (2021). URL https://www.architecture.com/knowledge-and-resources/resources-landing-page/riba-plan-of-work

[37] A. Arceo, M. Tham, G. Guven, H. L. MacLean, S. Saxe, Capturing variability in material intensity of single-family dwellings: A case study of Toronto, Canada, Resources, Conservation and Recycling 175 (2021) 105885.

[38] N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, K. Chobtham, A survey of bayesian network structure learning, Artificial Intelligence Review (2023) 1–94.

[39] K. Murphy, Bayes net toolbox for matlab: How to use the bayes net toolbox (2007). URL https://bayesnet.github.io/bnt/docs/usage.html

[40] M. Scutari, C. E. Graafland, J. M. Gutiérrez, Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms?, in: International Conference on Probabilistic Graphical Models, PMLR, 2018, pp. 416–427.

[41] J. Pearl, et al., Models, reasoning and inference, Cambridge, UK: CambridgeUniversityPress 19 (2) (2000) 3.

[42] C. H. Wagner, Simpson's paradox in real life, The American Statistician 36 (1) (1982) 46–48.

[43] BSI, BS EN 15978:2011 — Sustainability of construction works. Assessment of environmental performance of buildings. Calculation method (2011).

[44] G. Hammond, C. Jones, Embodied carbon, The inventory of carbon and energy (ICE). Version 3.0 (2019).

[45] S.-C. Lo, H.-w. Ma, S.-L. Lo, Quantifying and reducing uncertainty in life cycle assessment using the bayesian monte carlo method, Science of the total environment 340 (1-3) (2005) 23–33.

[46] E. Marsh, J. Orr, T. Ibell, Quantification of uncertainty in product stage embodied carbon calculations for buildings, Energy and Buildings 251 (2021) 111340.

[47] M. Huijbregts, et al., Uncertainty and variability in environmental life-cycle assessment, Citeseer, 2001.

[48] ISO, ISO 14025:2006 — Environmental labels and declarations. Type III environmental declarations, Principles and procedures (2006).

[49] O. P. Gibbons, J. Orr, C. Archer-Jones, W. Arnold, D. Green, How to calculate embodied carbon, Institution of Structural Engineers (IStructE), 2022.

[50] A. Ankan, J. Textor, pgmpy: A python toolkit for bayesian networks, arXiv preprint arXiv:2304.08639 (2023).

[51] B. D'Amico, Github repository for: Probabilistic inference of material intensities and embodied carbon in building structures (2023).
URL https://github.com/bernardinodamico/Probabilistic_carbon_building_structures

[52] S. Skiena, Hasse Diagrams (book chapter), in: Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Addison-Wesley, 1990, Ch. 5.4.2.

[53] C. De Wolf, E. Hoxha, A. Hollberg, C. Fivet, J. Ochsendorf, Database of embodied quantity outputs: Lowering material impacts through engineering, Journal of Architectural Engineering 26 (3) (2020) 04020016.

[54] B. D'Amico, R. J. Myers, J. Sykes, E. Voss, B. Cousins-Jenvey, W. Fawcett, S. Richardson, A. Kermani, F. Pomponi, Machine learning for sustainable structures: a call for data, in: Structures, Vol. 19, Elsevier, 2019, pp. 1–4.

[55] N. L. Zhang, D. Poole, A simple approach to bayesian network computations, in: Proc. of the Tenth Canadian Conference on Artificial Intelligence, 1994.

Appendix A.

Appendix A.1. Parent-child relationships among variables

Variable's parent(s) →	**Variable →**	**Variable's child(ren)**
Cladding type; Superstructure type	Masonry & Blockworks qty.	None
Superstructure type	Steel (sections) qty.	None
Superstructure type	Timber (products) qty.	None
Concrete elements; Foundations type	Reinforcement qty.	None

Foundations type;
Concrete elements;
Basement
	Concrete qty.	None

No. storeys; Superstructure unit-weight	Foundations type	Concrete type; Reinforcement qty.
Superstructure type	Superstructure unit-weight	Foundations type
Superstructure type	Concrete elem.	Reinforcement qty.
None	Basement	Concrete
None	Superstructure type	Masonry & Blockworks qty.; Timber (products) qty.; Steel (sections) qty.; Concrete elem.; Superstructure unit-weight
None	No. storeys	Superstructure unit-weight
None	Cladding	Masonry&Blockworks qty.
None	GIFA	Masonry & Blockworks qty.; Timber (products) qty.; Steel (sections) qty.; Concrete elem.; Superstructure unit-weight

Table A.1: Probabilistic dependencies encoded in the Directed Acyclic Graph shown in Figure 2.

Appendix A.2. Variable elimination

Variable Elimination is an algorithm to perform *exact* inference in probabilistic graphical models such as Bayesian Networks, first formalised by Zhang and Poole [55]. Considering the specific problem at hand: given a conditional probability query $P(Q|\mathbf{E})$, with $\mathbf{E} \subseteq \{S^*, C^{**}, N, F, B, G\}$ a subset of evidence variables (i.e. design choices) and $Q \in \{S, T, R, C, M\}$ a material quantity variable, the Variable Elimination algorithm involves an alternation of 'joining' *factors* ϕ (also termed *potentials*) and 'summing out' (or marginalise) variables. The initial factors are the CPTs encoding the Bayesian Network; with ref. to the variables' symbols in Table 2, initial factors are:

$$\begin{aligned}&P(S^*); P(C^{**}); P(N); P(B); P(G); P(T|S^*, G); P(S|S^*, G); P(S^{**}|S^*);\\&P(C^*|S^*); P(F|N, S^{**}); P(R|C^*, F, G); P(C|C^*, F, B, G); P(M|S^*, C^{**}, G)\end{aligned} \tag{A.1}$$

Assuming the query $P(C|s^*, n)$ to infer the probability distribution of *Concrete qty.* given *Superstructure type* $= s^*$ and *No. storeys* $= n$ as evidence, the relevant (unobserved) variables in the network are eliminated following a certain ordering, e.g for variables ordering B, F, C^*, S^{**}, R and G, the operations' order is:

1. join $\phi(B)$ and $\phi(C)$, then sum out B:

$$\begin{aligned}P(C, B|C^*, F, G) &= P(C|C^*, F, B, G)P(B)\\ \phi_1 = P(C|C^*, F, G) &= \sum_B P(C, B|C^*, F, G)\end{aligned} \tag{A.2}$$

2. join ϕ_1, $\phi(F)$ and $\phi(R)$, then sum out F:

$$\begin{aligned}P(C, F, R|C^*, n, S^{**}, G) &= P(C|C^*, F, G)P(F|n, S^{**})P(R|C^*, F)\\ \phi_2 = P(C, R|C^*, n, S^{**}, G) &= \sum_F P(C, F, R|C^*, n, S^{**}, G)\end{aligned} \tag{A.3}$$

3. join ϕ_2 and $\phi(C^*)$, then sum out C^*:

$$\begin{aligned}P(C, R, C^*|n, S^{**}, s^*, G) &= P(C, R|C^*, n, S^{**}, G)P(C^*|s^*)\\ \phi_3 = P(C, R|n, S^{**}, s^*, G) &= \sum_{C^*} P(C, R, C^*|n, S^{**}, s^*, G)\end{aligned} \tag{A.4}$$

4. join ϕ_3 and $\phi(S^{**})$, then sum out S^{**}:

$$\begin{aligned}P(C, R, S^{**}|n, s^*, G) &= P(C, R|n, S^{**}, s^*, G)P(S^{**}|s^*)\\ \phi_4 = P(C, R|n, s^*, G) &= \sum_{S^{**}} P(C, R, S^{**}|n, S^*, G)\end{aligned} \tag{A.5}$$

5. join ϕ_4 and $\phi(G)$, then sum out G:

$$\begin{aligned}P(C, R, G|n, s^*) &= P(C, R|n, s^*, G)P(G)\\ \phi_5 = P(C, R|n, s^*) &= \sum_G P(C, R, G|n, s^*)\end{aligned} \tag{A.6}$$

6. sum out R out of ϕ_5:

$$P(C|n, s^*) = \sum_R P(C, R|n, s^*) \tag{A.7}$$

Note how not all unobserved variables need to be eliminated. A variable is irrelevant for elimination if it is not an ancestor of the query variable or evidence variables, i.e. (T, S, C^{**} and M for the above query). Worth also of note: the elimination ordering has an impact on the algorithm's complexity, and while computing the optimal ordering that minimises the amount of arithmetic operations is an NP-hard problem, there are fairly efficient heuristics that can be used, such as *Min-fill*, that is, picking the variable at each elimination round that generates the smallest factor.

Appendix A.3. Embodied carbon coefficients

Figure A.1 shows the linear regression coefficients v_Q between material quantity values q, and corresponding embodied carbon values c_Q of data-points drawn from the P&M dataset:

$$c_Q = v_Q q \tag{A.8}$$

These slope coefficients represent the equivalent carbon factors v_Q of structural materials, from cradle to practical completion: $v_Q^{(A1-A5)} = v_Q^{(A1-A3)} + v_Q^{(A4-A5)}$ and were used in Eq.(13) here to derive the probability distribution of 'upfront' embodied carbon, $P(C_Q|Q)$, for each material variable Q. Values are: $v_{\text{Steel(Sections)}}$=1.5981; $v_{\text{Reinf.}}$=2.1293; v_{Concrete}=0.1298; $v_{\text{Timber(Prod.)}}$=0.5247; $v_{\text{Masonry\&Blockw.}}$=42.9874. For some materials like structural steel sections and steel reinforcement a perfect fit was found ($r^2 = 1.0$) whereas some variability can be seen for timber products and masonry & blockwork, presumably due to different (context specific) carbon coefficients used for different building project data-points reported in the dataset.

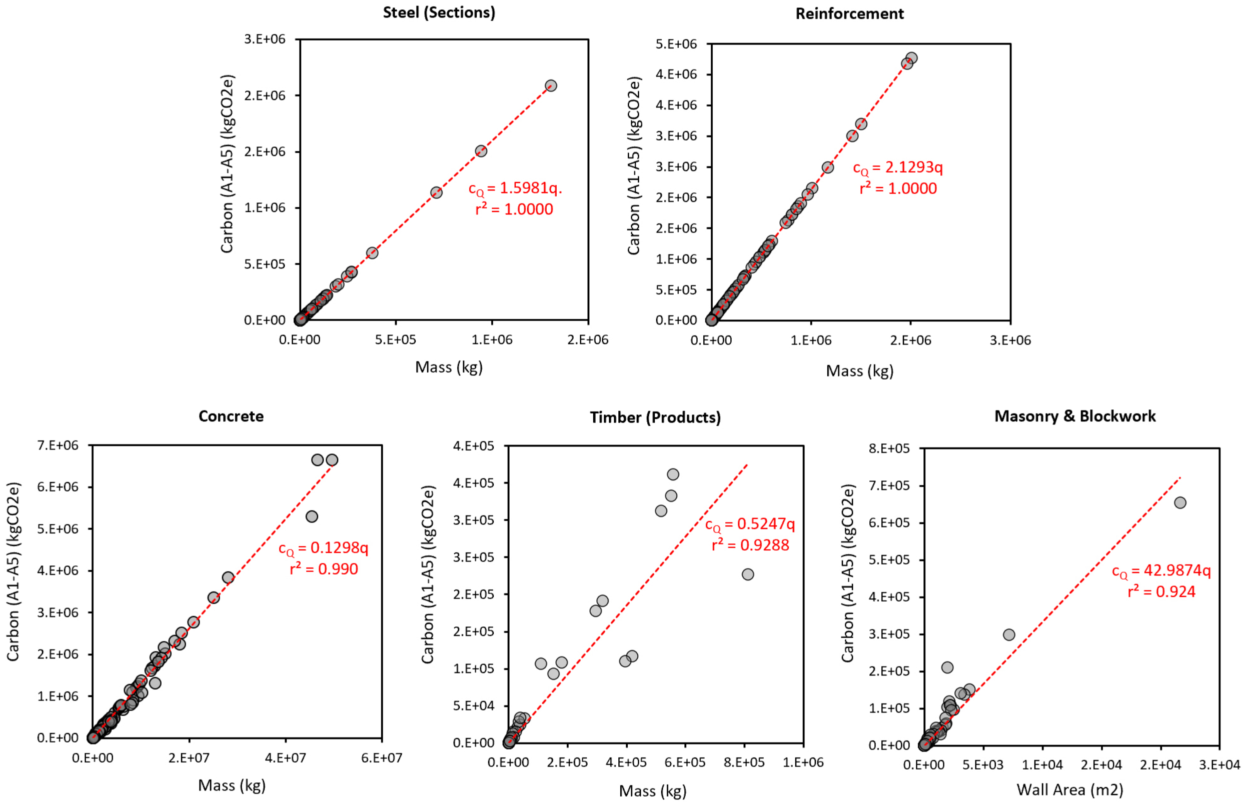

Figure A.1: Linear regression between material quantity variables and corresponding embodied carbon variables in the P&M dataset for life cycle stages A1-A5.

Appendix A.4. Probability distributions of material quantities

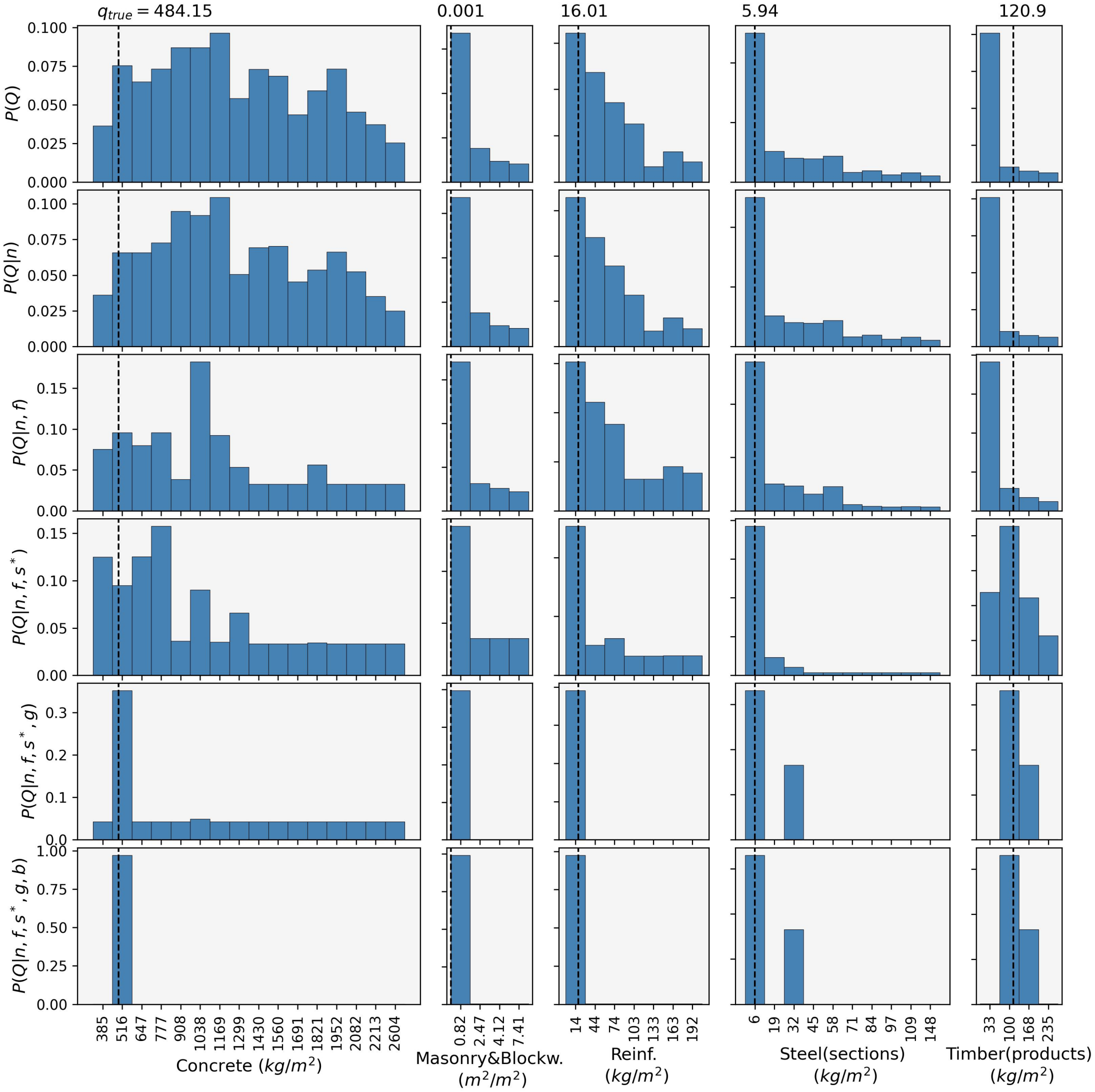

Figure A.2: Probability distributions of material quantities Q for the carbon queries shown in Figure 6. The top row shows inferred distributions given no evidence; the bottom row shows instead the same distributions given evidence of five design variable: N, F, S^*, G and B. The *true* value for each material quantity, q_{true}, is reported at the top of each column. Note: ticks values on the horizontal axes indicate the mid-value of the bins —ranges for each bin are reported in Table 2.