A probabilistic building characterization method for district energy simulations

De Jaeger, Ina; Lago, Jesus; Saelens, Dirk

DOI
10.1016/j.enbuild.2020.110566

Publication date
2021

Document Version
Final published version

Published in
Energy and Buildings

Energy & Buildings 230 (2021) 110566

Contents lists available at ScienceDirect

Energy & Buildings

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

A probabilistic building characterization method for district energy simulations

Ina De Jaeger [a,b,c,*], Jesus Lago [a,b,d], Dirk Saelens [a,c]

[a] EnergyVille, Thor Park 8310, BE-3600 Genk, Belgium
[b] Flemish Institute for Technological Research (VITO), Boeretang 200, BE-2400 Mol, Belgium
[c] KU Leuven, Department of Civil Engineering, Building Physics Section, Kasteelpark Arenberg 40 – box 2447, BE-3001 Leuven, Belgium
[d] Delft University of Technology, Delft Center for Systems and Control, Mekelweg 2, NL-2628CD Delft, Netherlands

ARTICLE INFO

Article history:
Received 26 February 2020
Revised 6 October 2020
Accepted 13 October 2020
Available online 16 October 2020

Keywords:
Urban building energy modelling
District energy simulation
Input data
Uncertainty
Multivariate probability distribution
Scenario generation

ABSTRACT

To assess the impact of implementing energy efficiency and renewable energy measures, urban building energy models are emerging. In these models, due to the lack of data, the natural variability of the existing building stock is often highly underestimated and uncertainty on the simulated energy use arises. Therefore, this work proposes a probabilistic building characterization method to model the variability of the existing residential building stock. The method estimates realistic distributions of five input variables: U-values of the floor, external walls, windows and roof as well as window-to-wall ratio, based on known data (location, geometry and construction year). First, quantile regression has been implemented to generate the uncorrelated distributions based on the Flemish energy performance certificates database. The accuracy of the marginal distributions is good, as the empirical coverage on the 50%, 80%, 90% and 98% prediction interval deviates 0.6% at most. However, it is needed to include the correlations between these variables. Hence, three main methods to build multivariate distributions from marginal distributions and to draw correlated samples are implemented and extensively compared. The Gaussian copula method is put forward as the preferred method. Considering the mean-maximum discrepancy (MMD), this method performs eight times better than the uncorrelated case (MMD of 0.0027 versus 0.0228).

1. Introduction

1.1. Research background

To lower the environmental impact of our existing building stock and to mitigate climate change, both increasing the energy efficiency and integrating renewable energy sources are measures to be pursued. As an example, Le Guen et al. studied how to combine both kind of measures to improve the energy sustainability of a Swiss village [1]. Orehounig et al. optimised the share of energy demand covered by renewables through integrating decentralized energy systems in a Swiss village [2]. Hirvonen et al. assessed the cost-effectiveness of apartment buildings renovation measures, including both insulating and replacing heating, ventilation and air conditioning (HVAC) systems on building level [3].

A lot of recent research initiatives study increasing energy efficiency and integrating renewable energy sources on a district or city level to include the synergy effects that result from the heterogeneity of the existing building stock. As an example, Lawrence et al. focused on integrating smart building in the electrical grid [4], while Kensby et al. analysed the potential of buildings to be used as thermal energy storage in district heating systems [5]. In this context, urban building energy models (UBEMs) are emerging and are used to quantify the operational building energy use on district or city level through building-by-building simulation [6]. UBEMs are typically bottom-up building physical models [7], enabling to analyse the current status of the building stock and to assess possible future scenarios that combine energy efficiency measures with renewable energy integration. They allow for studies on multiple levels, from street to district to city level.

Abbreviations: UBEM, urban building energy model; HVAC, heating, ventilation and air conditioning; GIS, geographical information system; WWR, window-to-wall ratio; EPC, energy performance certificate; QR, quantile regression; OLS, ordinary least squares; CDF, cumulative distribution function; MMD, mean-maximum discrepancy.

* Corresponding author at: EnergyVille, Thor Park 8310, BE-3600 Genk, Belgium.
E-mail address: ina.dejaeger@kuleuven.be (I. De Jaeger).

https://doi.org/10.1016/j.enbuild.2020.110566
0378-7788/

1.2. Research gap

However, rather than fully characterising the whole district or city, UBEMs usually make use of *archetype buildings* due to the lack of sufficient input data on building level [6], reducing the actual variability of the existing building stock. An archetype building is a representative building for a group of similar buildings. More information on archetype buildings can be found in De Jaeger et al. [8].

Although UBEM simulations have been reported to correspond reasonably well to measured energy use data on higher aggregation levels (city to nation), with errors of the considered studies ranging from 7 to 21% [6], the errors increase significantly when focusing on smaller scale examples. Particularly for analysis on smaller scales (~100 dwellings), the use of building archetypes can be questioned. These analyses include amongst others the study of low voltage grids [9], the exploration of demand-side management in electrical grids [10], the design and operation of distributed multi-energy systems [11], the design and operation of district heating networks [12], the trade-off between improving in energy efficiency and employing renewable energy within districts and cities [13].

Two examples of archetypes performing worse on smaller scales can be found in Orehounig et al. [14] and De Jaeger et al. [15]. Orehounig et al. [14] modelled a Swiss village of 100 buildings using both a simplified (i.e. archetype) and a detailed (i.e. building-by-building) modelling approach and compared the simulation results against measured energy use. On municipality level, an 8% deviation in annual energy demand between the simulations and the measurements is found for the detailed approach. However, the estimation for a smaller group of buildings or for a single building can be significantly worse. In their study, the estimation for buildings constructed before 1980 deviates approximately 50% deviation in annual energy demand for the simplified approach, which makes use of archetypes. Deviations for estimations on building level are even higher. De Jaeger et al. [15] compared the use of Belgian TABULA archetypes [16] to the use of geospatial data for a small district of 99 buildings. They concluded that, depending on the building typology (detached, semi-detached or terraced dwelling), the TABULA archetypes underestimate the peak heat demand by 26% to 95% on average and the annual heat demand by 3% to 80% on average. In other words, the geometry of the TABULA archetypes is not representative for this particular district. Moreover, the archetype approach fails to include the non-negligible variability in building geometry that is characteristic for the existing building stock.

These increasing errors are caused by the two tasks that are performed to define archetypes, more in particular *segmentation* and *characterization* [6]. First, the building stock is *segmented* or divided into groups of similar buildings, e.g. based on their building type and construction year. Segmentation inherently reduces the variability of the existing building stock within UBEMs. In other words, the actual variability within the existing building stock is overly simplified (e.g. all buildings are categorised based on their construction year, but possible renovations are often not considered). However, acknowledging this variability is amongst others crucial for the optimal design of district energy systems [17]. Second, the representative building for each group is *characterized* or defined. As shown by the examples above, it should be ensured that the archetype buildings are representative. Ensuring representative archetypes has successfully been the focus of Ghiassi and Mahdavi [18], Tardioli et al. [19] and De Jaeger et al. [8]. In addition, *measurement data* can be obtained from distribution system operators and can be used both to calibrate the model [20] and to estimate the simulation error.

Unfortunately, the issue of underestimating the variability can only be tackled by shifting away from archetypes and characterising each building separately. To the best of the authors' knowledge, there is no available method that characterises existing districts without using energy performance or building archetypes and thus no available method that includes the full variability of the existing districts.

1.3. Research objectives

To fill this gap, this paper presents a probabilistic building characterisation method. By using this method, every building of the UBEM is characterised by a particular probability density function for the U-values of the ground floor, external walls, windows and roof as well as the window-to-wall ratio (WWR) based on known data – i.e. construction year, building location and building geometry. As a result, correlated samples for the U-values and the WWR can be obtained on building level. The probabilistic approach does not only allow to estimate an average value per building, but also to include the probability of being renovated.

To characterise each building separately, a considerable number of input parameters is required for each building. These include location and geometry, thermal quality of the building envelope, HVAC systems, renewable energy systems, building appliances and occupant behaviour. Although these characteristics could be acquired per building through on-site measurements or surveys, the data acquisition effort becomes infeasible on district or city level, as discussed by Hong et al. [21] and Monteiro et al. [22]. As an alternative to real data, *statistical data* on the building envelope and system characteristics can often be obtained from governmental databases such as the energy performance certificates (EPC) databases in Europe [23]. EPCs are labels that inform consumers of the energy efficiency of buildings they plan to purchase or rent. The Flemish EPC database is therefore a valuable resource for energy performance-related data of buildings (i.e. building type, construction year, building geometry, thermal performance of the building envelope, information on the HVAC systems, ...). However, privacy issues are often the key argument for not sharing the data on building level. In addition, the data quality of these databases should be treated with care, as discussed more elaborately in Section 2.1.

In this work, the Flemish EPC database is employed to obtain the probability density functions for the building energy related data and to relate these to parameters that are known on individual building level through geographical information system (GIS) and cadastral data – i.e. construction year, building location and building geometry. As the different building energy related parameters appear to be correlated, three methods to draw correlated samples from these marginal – i.e. uncorrelated – probability distributions are implemented in multiple variants and extensively compared. Including the correlations is of utmost importance to achieve realistic samples. Based on four numerical performance indicators, the Gaussian copula method is selected to be included in the probabilistic building characterization method. In addition, although this method is harder to implement, it is easier in use.

The novelty of this method is the ability to characterise each building of a UBEM separately in a probabilistic way. This method is also particularly interesting to obtain realistic input parameter variations to perform uncertainty and sensitivity analyses of the energy demand for existing residential neighbourhoods within future work. As these simulations are often used to make decisions towards a more sustainable city or district, it is highly important to include the impact of uncertainties [24–26]. More particular use cases are listed in Section 4.

Although this probabilistic approach can be extended to the building energy systems as well as to other building typologies

(e.g. office buildings), a first assessment of the usability of the EPC database focuses solely on the building envelope characteristics of Flemish single-family dwellings. In this work, only continuous building-level variables, for which data were available, are discussed. The probabilistic method will be extended to the building energy systems within future work. The building energy systems are categorical variables and require a classification method instead of a continuous method. In addition to the building-level parameters, local climate conditions and urban-level parameters (e.g. morphology) should be included in the UBEM [27], but this is also considered to be future work.

In the next Section, the probabilistic methodology to allocate building energy related data in UBEMs is introduced. Three main methods to characterize the multivariate probability distribution of the U-values and the WWR and multiple variants of these methods are introduced. Then, these distributions are used to generate correlated samples. Additionally, four performance indicators to compare these methods are described. In Section 3, the performance indicators for the three methods are presented and discussed. Finally, in Section 4, the conclusions are drawn.

2. Methodology

In this Section, the workflow of the probabilistic building characterisation method is presented, which is also illustrated in Fig. 1. First, the marginal distributions for the U-values and the WWR are obtained, through quantile regression (QR). Second, the multivariate distributions of the U-values and the WWR are determined based on the marginal distributions and samples are drawn. To build the multivariate distributions and to draw correlated samples, three main methods in multiple variants are proposed. The three main methods are the sequential method, the Gaussian copula method and the empirical copula method. The most appropriate method is identified based on four performance indicators and will be included in the probabilistic building characterisation method. After presenting quantile regression to obtain the marginal distributions and describing the different methods to build the multivariate distributions, the performance indicators are introduced.

2.1. Marginal distributions for the U-values and WWR

In a first stage, the marginal distributions for the building envelope properties (i.e. the U-values of the floor, external walls, windows and roof as well as the WWR) are obtained through quantile regression, based on available data (i.e. building geometry and construction year). A marginal distribution of a random variable X (e.g. U-value of the roof) includes the probabilities of all possible values for X, regardless of the values of other correlated variables (e.g. U-value of the external walls and the windows). Therefore, marginal distributions do not consider any correlation. This subsection first describes the available data in Flanders as well as the Flemish EPC database, which is used to fit the quantile regression models. Then, the theory of quantile regression is introduced. Finally, the generation of the marginal distributions in practice is explained.

For Flanders, the available data for all buildings consist of construction year, building location and building geometry. Building geometry data can be obtained from the Flemish GIS, but for this study, they are obtained from a CityGML model of the city of Genk (Belgium) with level of detail (LOD) 2 [28]. The available building geometry and location data include postal code, building type (terraced, semi-detached or detached dwelling), building volume, building height, ground floor area, façade area and roof area. The heated floor area can be deduced from an assumed number of storeys. In this work, the assumed number of storeys is defined as the maximum number of floors with a height of at least three metres that fit within the ridge height [15]. As many single-family dwellings are characterised by building extensions with a lower height than the main volume, the heated floor area is calculated for main buildings and building extensions separately to avoid an overestimation. Finally, the construction year can also be obtained from the Flemish cadastral database, which is a land registry that contains, among others, information on the ownership, land use, building geometry, and building construction year for taxation purposes.

To obtain the marginal probability distribution function for all U-values and the WWR of the buildings, quantile regression models [29] are built with data from the Flemish EPC database as input. Since EPCs are labels that report on the energy efficiency of buildings, a significant amount of useful data is available from the Flemish EPC database. These data include building type, construction year, building geometry, thermal performance of the building envelope as well as information on the HVAC systems and provide an essential link between the known and the unknown data on existing districts. However, statistical methods are required as the data cannot be shared on building level due to privacy issues. In addition, not all buildings are included in the database. In this work, an anonymised version of the EPC database that excludes address-related information is used, although the municipality in which the dwelling is situated is known.

It is important to note that the use of the EPC dataset to determine distributions on the U-values and the WWR should be considered with care because of two aspects. First, EPCs are only established before purchasing or renting a building and buildings are very likely to be renovated immediately after they have been purchased. Second, only EPCs of existing buildings are considered

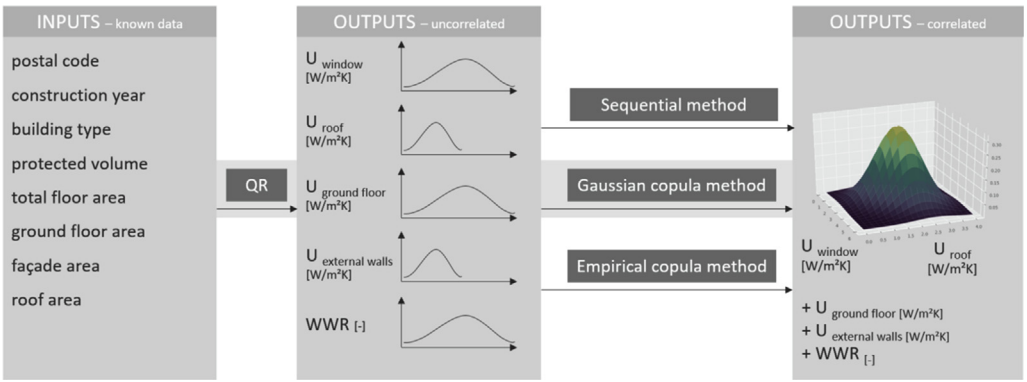

Fig. 1. Graphical overview of the methodology, as used in this work.

and their exact construction layers and materials are often unknown. When this is the case, conservative values are used by default, which are also visible in Fig. 2. Fig. 2 illustrates the U-value of the external wall and the roof for the buildings of the EPC dataset that were constructed in 1900. In the scatter plot, multiple vertical (U-value of 1.7 and 2.7 W/m^{2}K) and horizontal lines (U-value of 2.1 and 2.9 W/m^{2}K) are clearly visible, indicating often applied default values for the U-values of the external wall and the roof respectively. As a result, the EPC dataset gives a rather conservative view of the existing building stock. These drawbacks of using EPC data highlight the need for more accurate data on the current state of our existing building stock. Despite these shortcomings, the EPC dataset is used as it is the best data source that is currently available.

Quantile regression, introduced by Koenker and Bassett [29], estimates a model of the quantiles of the conditional distribution of the response variable as functions of observed covariates. This can be compared with an Ordinary Least Squares (OLS) method in which the conditional mean is estimated by minimizing the squared residuals. In particular, instead of estimating the mean, QR models estimate the conditional distribution of the response variable. In more detail, in OLS, the sample mean μ of a variable y, which is an estimate of the unconditional population mean E(Y) based on n data points, is found by solving the following problem:

$$\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \tag{1}$$

Likewise, an estimate of the conditional expectation function E($Y|x$) can be equally found by OLS by replacing μ by a parametric function $\mu(x, \beta)$:

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2 \tag{2}$$

In QR, the unconditional τ^{th} quantile of y, i.e. q_τ, can be found by solving the following problem:

$$\min_{q_\tau\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - q_\tau) \tag{3}$$

where $\rho_\tau(y_i - q_\tau) = \tau^*(y_i - q_\tau)$ for$(y_i - q_\tau) > 0$ and $\rho_\tau\,(y_i - q_\tau) = (\tau-1)^*(y_i - q_\tau)$ for $(y_i - q_\tau) < 0$. Similarly, an estimate of the conditional τ^{th} quantile of y can be found by replacing q_τ by a parametric function $q_\tau(x_i, \beta)$:

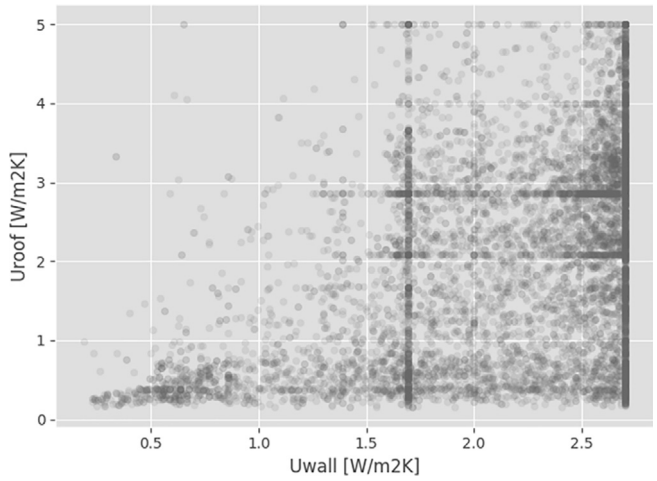

Fig. 2. The U-values of the external wall and the roof for the buildings of the EPC dataset that were constructed in 1900. This Figure shows a correlation between the U-values of the external wall and the roof and shows the importance of including correlations. Also, the default values of the EPC dataset are visible. This is an inherent shortcoming of the employed dataset.

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - q_\tau(x_i, \beta)) \tag{4}$$

By definition, quantiles are in fact a representation of the cumulative distribution functions (CDF) of the corresponding random variable. More in particular, given the CDF F(y) of a random variable y, q_τ, i.e. the τ^{th} quantile, relates to F(y) as follows: F$(q_\tau) = \tau$. Therefore, QR models are able to characterize a complete range of quantiles and thus approximate the full CDF of y. The QR models in this work are linear and the optimisation problem to estimate the quantiles is thus solved through linear programming [29].

To setup the marginal distributions, an anonymised version of the EPC database, the StatsModels and the scikit-learn Python packages were used. Before the EPC data could be used to fit the QR models, some invalid data points are removed, as they contain incorrect information. This has been done by visually observing scatter plots and applying filters to the data. For construction year, total floor area, protected volume, only values between the 1st and the 99th percentile were kept. For the external wall area, only values between the 2.5th and the 97.5th percentile were kept. For all other geometrical parameters, only values between the 2.5th and the 99th percentile were kept. The U-value of the external walls is between 0.15 and 3 W/m^{2}K. The U-value of the windows is between 0.8 and 6 W/m^{2}K. The U-value of the ground floor is between 0.15 and 3 W/m^{2}K. The U-value of the roof is between 0.15 and 6 W/m^{2}K. By using an 80/20 ratio for training and test dataset, the QR models were built based on a training dataset of 340,618 dwellings in the EPC database, leaving 85,155 dwellings as an out-of-sample test dataset. The accuracy of the QR models has not been tested for less than 340,618 data points nor for a different ratio of training and test dataset, although this could be explored in future work.

In this work, postal code, building type, construction year, total floor area, protected volume, ground floor area, façade area (opaque plus transparent) and roof area were considered as explanatory variables, since they are available for all buildings and a preliminary analysis showed their relevance. Then, QR models are fitted for each output variable (U-values of the ground floor, external walls, windows and roof as well as WWR) and each τ^{th} quantile, with τ ranging from 0.01 to 0.99. As a result, 495 models are fitted. Subsequently, to generate the marginal distributions for the test dataset buildings, the explanatory variables for each dwelling are fed into each of the QR models. In other words, the value corresponding to each τ^{th} quantile is predicted for each output variable for each dwelling. That way, the CDFs for the five output variables are characterized by aggregating the QR models for each τ^{th} quantile.

2.2. Correlated samples for the U-values and WWR

While marginal distributions of random variables can be used to generate samples, these samples are only realistic if the random variables are completely independent in the multivariate case, i.e. multiple random variables. In other words, if the random variables are correlated and samples are generated by solely using their marginal distributions, the generated values would be uncorrelated and would not represent realistic samples. In this case, the variables are not independent and their correlations are clearly visible in the EPC dataset, as illustrated in Fig. 2. Fig. 2 shows that buildings with a *good* external wall are more likely to have a *good* roof compared against the buildings with a *bad* external wall. Therefore, after obtaining the marginal distributions for the different variables, it is important to infer their multivariate relations (i.e. their multivariate distribution or the correlations between the different variables), so that realistic samples can be generated.

In this work, three main methods to build multivariate distributions from marginal distributions and to draw correlated samples are proposed and extensively compared: the sequential method, the empirical copula method and the Gaussian copula method. The three main methods all envisage another way to include the correlations between the U-values and the WWR. The *sequential method* simply changes the input variables that are used to fit the QR models, whereas the *empirical copula method* and the *Gaussian copula method* both employ a particular copula, i.e. a function that link multivariate distributions to their univariate marginal distributions. As illustrated in Fig. 3, different variations of the three methods are tested. The three methods and their variations are introduced more elaborately below. First, it is presented how these methods are used to generate correlated samples based on the test dataset. Then, it is explained how these methods are implemented based on the training dataset.

2.2.1. Sequential method

The first method is developed within this paper and is referred to as the *sequential method or SM* (Fig. 3). The method is easy to implement, as it sequentially fits different QR models. The main idea is to sequentially build QR models where successive QR models are based on both the original input variables and the previously estimated output variables, instead of fitting the QR models separately for all output variables based on an identical set of input variables (i.e. the marginal distributions). In this way, the generated samples are correlated, as the previously estimated output variables are taken into account to estimate the next output variable. First, it is explained how to generate samples, then it is described how to implement the method.

Generating correlated samples using this method is rather straightforward. First, the probability distribution for the first output variable is predicted based on the initial input data. Then, a particular value is sampled from this distribution. Subsequently, the probability distribution for the second output variable can be predicted based on the initial input data and the sampled value for the first output variable, as the QR model of the second variable uses the first random variable as input. Again, a particular value is sampled from this distribution. This is repeated until a value is sampled for the last output variable.

The method is implemented based on the training dataset in different steps. First, the order in which the output variables will be sampled is defined. Then, an estimate of the conditional expectation function of the first output variable Y_1, $E(Y_1|x)$, is calculated for each of the 99 quantiles based on the input variables x, using Equation (4). In other words, the QR model for the first output variable is fitted based on the original set of input variables. Next, an estimate of the conditional expectation function of the second output variable Y_2, $E(Y_2|x,Y_1)$, is calculated for each of the 99 quantiles based on the input variables x and the first output variable Y_1. In other words, the QR model for the second output variable is fitted based on the original set of input variables plus the first output variable. Subsequently, the QR model for the third output variable is fitted based on the original set of input variables plus the first two output variables. This is repeated until the QR model for the last output variable is fitted.

To find the most optimal order in which the output variables should be sampled from the EPC dataset, the order was perturbed in a preliminary analysis, resulting in 120 combinations. This preliminary analysis showed that the different orders performed rather similar. Therefore, and due to the rather poor performance of the method (Section 3.2.1), only seven variants are shown in this work and are listed in Table 1.

2.2.2. Empirical copula method

The second method is proposed by Clark et al. [30] and is referred to as the *Empirical copula method or ECM* (Fig. 3). This method represents a simple methodology to define the copula function based on historical data. It has been applied first to reconstruct space–time variability in joint forecasts of precipitation and temperature [30] and has been implemented in this work to characterize the energy-performance related parameters of existing residential buildings (i.e. joint scenarios for the U-values and the WWR).

The main idea of the method is to use historical data to build rank correlations [31]. The correlation between the random variables is thus modelled based on their ranks. First, for each variable X, i.e. U-values of the ground floor, external walls, windows and roof and WWR, the method uses its marginal distributions and generates N uncorrelated samples:

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_N) \tag{5}$$

Then, these N samples are sorted by value from small to large, resulting in $\boldsymbol{\chi}$ (Equation *(6)*).

$$\boldsymbol{\chi} = \left(x_{(1)}, x_{(2)}, \cdots, x_{(N)}\right),\ x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)} \tag{6}$$

These ranked uncorrelated samples are illustrated in the middle table of Fig. 4. $\boldsymbol{\chi}$ represents a single column in this table. Subsequently, a template Z is used to define the ranks of historical values based on historical data. In Z, each column corresponds to a particular output variable and contains n buildings from a historical dataset, given by $\boldsymbol{Y}$ (Equation *(7)*). This vector $\boldsymbol{Y}$ could also be sorted by value from small to large, resulting $\boldsymbol{\gamma}$ (Equation *(8)*).

$$\boldsymbol{Y} = (y_1, y_2, \cdots, y_N) \tag{7}$$

$$\boldsymbol{\gamma} = \left(y_{(1)}, y_{(2)}, \cdots, y_{(N)}\right), y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)} \tag{8}$$

As a result, each row corresponds with a particular data point, i.e. building, and indicates the relative rank of each of the five variables of the particular data point. Now, consider vector $\boldsymbol{Z}$ that contains the indices describing the original observation numbers $1, 2, \cdots, N$ as the values in the ordered vector $\boldsymbol{\gamma}$ appeared in $\boldsymbol{Y}$. This vector $\boldsymbol{Z}$ corresponds to the particular column in the ranked template Z that is also illustrated in the left table of Fig. 4. As an example, in this illustration, the WWR of the first building selected from the historical dataset is the highest of all selected buildings, the WWR of the second selected building is second lowest of all selected buildings and so on. In other words, the values of the N

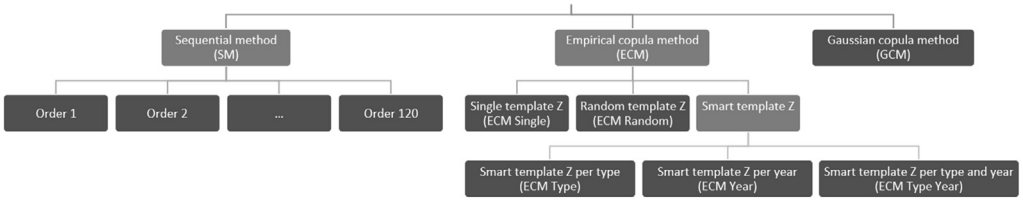

Fig. 3. Overview of the considered methods to draw correlated samples. Three main methods are considered: the sequential method, the empirical copula method and the Gaussian copula method. For the sequential method and the empirical copula method, different variations have been evaluated.

Table 1
Overview of the different orders that are used in the different variations for the sequential method in this work.

Name	Order (First> second > third > fourth > fifth)
SM 1	$U_{floor} > U_{wall} > U_{roof} > U_{window} > WWR$
SM 2	$U_{floor} > U_{wall} > U_{roof} > WWR > U_{window}$
SM 3	$U_{floor} > U_{wall} > U_{window} > U_{roof} > WWR$
SM 4	$U_{floor} > U_{wall} > U_{window} > WWR > U_{roof}$
SM 5	$U_{floor} > U_{wall} > WWR > U_{roof} > U_{window}$
SM 6	$U_{floor} > U_{wall} > WWR > U_{window} > U_{roof}$
SM 7	$U_{floor} > U_{window} > U_{wall} > U_{roof} > WWR$

buildings in template Z are simply replaced by their relative ranks. This matrix of ranks is used as the template Z representing the rank variable correlation.

Using the template Z, the ranked uncorrelated samples are combined to mimic the ranks in this template. In other words, the correlated sample values are constructed, represented by the reordered vector $\boldsymbol{X^{SS}}$, following Equation *(9)*:

$$\boldsymbol{X^{SS}} = (x_1^{ss}, x_2^{ss}, \cdots, x_N^{ss}), where \tag{9}$$

$$x_q^{ss} = x_{(r)} \tag{10}$$

$$q = \boldsymbol{Z}[r] \tag{11}$$

$$r = 1, 2, \cdots, N \tag{12}$$

These correlated samples are also illustrated in the right table of Fig. 4. In this illustration, the ranked template Z prescribes that the smallest value of the WWR, $x_{(1)} = 0.05$, should be placed in correlated sample $q = \boldsymbol{Z}[1] = 5$. Therefore, $x_5^{ss} = x_{(1)} = 0.05$. The ranked template Z also prescribes that the second smallest value of the WWR, $x_{(2)} = 0.1$, should be placed in correlated sample $q = \boldsymbol{Z}[2] = 2$. Therefore, $x_2^{ss} = x_{(2)} = 0.1$ and so on. If N correlated samples of the five variables are to be generated, then template Z should be an Nx5 matrix and thus contain N buildings from the historical dataset.

In this work, multiple definitions of the template Z are proposed. A first definition of template Z is to use one single template to draw the N samples for all buildings of the test dataset. This method is referred to as *ECM Single*. Obviously, the results then highly depend on the selected N buildings from the training dataset. In a second approach, referred to as *ECM Random*, there are as many templates Z as there are buildings in the test dataset. Every template Z is based on N randomly selected dwellings. As a third approach, the template Z is defined by drawing samples from a subset of similar buildings. Three definitions have been used within this third approach. First, the template Z for a particular building is based on buildings with the same building type (i.e. terraced, semi-detached or detached). This method is referred to as the *ECM Type*. Second, the template Z for a particular building is based on buildings that were constructed five years before or after the particular building. This method is referred to as the *ECM Year*. Third, the template Z for a particular building is based on buildings that have an identical building type and that were constructed five years before or after the particular building. This method is referred to as the *ECM Type Year*.

2.2.3. Gaussian copula method

The third method, originally developed by Pinson et al. [32], is referred to as the *Gaussian copula method (GCM)* (Fig. 3). This method infers a parametric multivariate Gaussian copula using the marginal distributions of the random variables. This parametric multivariate Gaussian copula is characterized by a correlation matrix R, which is estimated based on the training dataset and is used to generate samples. It has been applied first to create statistical scenarios for short-term wind power production [32] and has been implemented in this work to create statistical scenarios for the U-values and the WWR of existing residential buildings.

To obtain the multivariate Gaussian copula, the key is to transform the original output variables of the buildings in the training dataset into a multivariate Gaussian random variable, of which the interdependence structure can be summarized by a unique covariance matrix. First, the marginal distributions for all output variables for all buildings of the training dataset are generated (i.e. the values that correspond to each of the 99 percentiles). Then, it is determined in which percentile of this estimated marginal distribution the real value is situated for all output variables. Given the CDF $F(x_k)$ of a random variable x_k, $q_{\tau,k}$, i.e. the τ^{th} quantile for variable x_k, relates to $F(x_k)$ following Equation *(13)*:

$$\tau_k = F(q_{\tau,k}) \sim \text{Uniform}[0, 1] \tag{13}$$

Based on that, each historical value x_k is mapped to its quantile τ_k. By doing so, the historical values are mapped to a set of uniformly distributed values in the interval [0,1]. Then, given this uniformly distributed τ_k, a second transformation is performed to obtain a Gaussian variable with zero as a mean and one as a standard deviation, following Equation *(14)*:

$$\Phi^{-1} : x \rightarrow \sqrt{2} erf^{-1}(2x - 1) \tag{14}$$

where erf^{-1} is the inverse error function and Φ^{-1} is the probit function that is equal to the inverse of the Gaussian CDF. Hence, applying the probit function to the uniformly distributed variable τ_k, as shown in Equation *(15)*, results in Y_k, which is a random variable that Gaussian distributed with zero as a mean and one as a standard deviation.

$$Y_k = \Phi^{-1}(\tau_k) \sim Normal(0, 1) \tag{15}$$

Finally, the method assumes that the random vector $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_5)$ containing the transformed random variables Y_k for each of the five output variables is a multivariate normal distribution with 0 as a mean and the desired covariance matrix as

Ranked template Z

$U_{ground\ floor}$	U_{wall}	U_{window}	U_{roof}	WWR
1	1	1	1	5
2	2	4	3	2
3	5	5	5	4
4	4	2	2	3
5	3	3	4	1

Ranked uncorrelated samples for real building

$U_{ground\ floor}$	U_{wall}	U_{window}	U_{roof}	WWR
0.3	0.5	1.5	0.2	0.05
0.8	1.0	2.7	0.5	0.1
1.0	1.2	3.5	0.7	0.15
2.0	1.4	4.0	1.0	0.2
2.5	2.0	5.9	1.5	0.25

Correlated samples for real building

$U_{ground\ floor}$	U_{wall}	U_{window}	U_{roof}	WWR
0.3	0.5	1.5	0.2	0.25
0.8	1.0	4.0	0.7	0.1
1.0	2.0	2.9	1.5	0.2
2.0	1.4	2.7	0.5	0.15
2.5	1.2	3.5	1.0	0.05

Fig. 4. Simplified illustration of how to obtain correlated samples using the empirical copula method. In this example, five correlated samples are drawn. Each row in the template Z represents a dwelling included in the historical dataset. In template Z, all U-values and WWRs are replaced by their ranks. To create five correlated samples for a real dwelling, five uncorrelated samples are generated first and ranked from small to large. Template Z describes which rank to combine with which in the newly generated uncorrelated samples to obtain correlated samples.

covariance. From a multivariate Gaussian variable, the covariance matrix can easily be determined. This covariance matrix only needs to be calculated once based on the training dataset. When generating correlated samples, only the covariance matrix is required, but the training data is not.

The main idea of the method is thus to calculate the covariance matrix based on historical data and use this matrix to include the correlations in new samples. After estimating the multivariate Gaussian copula, the procedure to generate samples is illustrated in Fig. 5. To obtain the correlated samples, two transformations, opposite to Equations *(15) and (13)*, are performed. First, a correlated sample is drawn from a multivariate normal distribution characterised by a 0-mean and the defined covariance matrix, which has 1-values on its diagonal (i.e. a unit standard deviation). Then, these values are transformed to their percentiles. This transformation, described by Equation *(16)*, is opposite to the transformation described by Equation *(15)*.

$$\tau_k = \Phi(Y_k) \tag{16}$$

Finally, these percentiles are transformed to the real values for all output variables based on their marginal probability distributions as generated by the QR models. Again, this transformation, described by Equation *(17)*, is the inverse of the transformation described by Equation *(13)*.

$$q_{\tau,k} = \mathrm{F}^{-1}(\tau_k) \tag{17}$$

This way, the value of variable k, $q_{\tau,k}$, can be obtained, for the 5 variables and for the desired number of samples.

2.3. Performance indicators

This work proposes a probabilistic building characterisation method to enrich the available data with energy performance-related data within UBEMs. Five parameters are considered, i.e. the U-values of the floor, external walls, windows and roof as well as the WWR. First, QR has been proposed to generate the marginal distributions for these five variables. Then, the different methods to build multivariate distributions from marginal distributions and to draw correlated samples have been presented. In this work, 12 random samples are generated for all 85,155 buildings of the test dataset. The different implementations of the probabilistic building characterisation method are extensively compared based on multiple performance indicators. These performance indicators are introduced below.

To check the accuracy of these marginal distributions, the *empirical coverage* is checked at different prediction intervals. In this work, the 50%, 80%, 90% and 98% prediction intervals are considered and the empirical coverage on these intervals is computed. As an example, the 90% prediction interval is discussed. For the 90% prediction interval, the empirical coverage is equal to the percentage of all buildings in the test dataset of which the real value falls within the predicted 5th and 95th quantiles and should ideally be close to the theoretical range of 90%. The empirical coverages are shown in the next Section.

To check the accuracy of the multivariate distributions, the ideal performance indicator measures the distance between the real multivariate distribution of all 85,155 buildings in the test dataset and the generated multivariate distributions of all 85,155 buildings in the test dataset. This is exactly achieved by the main performance indicator, the *mean-maximum discrepancy* (MMD) [33]. The MMD is a distance on the space of probability measures that is used to compare statistical distributions and that has been used in several machine learning applications and non-parametric testing. Formally, this distance is defined as the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilberspace. The MMD is a good and complete performance indicator, as it quantifies the distance between the real multivariate probability distribution and the multivariate probability distributions generated by the different methods. The MMD allows to assess the relative performance of the different methods compared to each other. However, it is not straightforward to correctly interpret the magnitude of the MMD for the best performing method. In other words, although the MDD is small for the best performing method, it is difficult to assess whether this method is representative for the real probability distribution based on MDD. Therefore, two additional performance indicators are proposed in this work.

The first additional performance indicator focuses on the correlation between the U-values, in sets of two, and is therefore referred to as the *correlation error*. To calculate the correlation error, all U-values (i.e. of the floor, external walls, windows and roof) are first labelled as *bad*, *moderate* and *good*. The thresholds between these three categories are defined based on the 33th and the 66th percentile of the real values of the buildings in the test dataset respectively. For every set of U-values (e.g. U-value roof and U-value external wall), the percentage of buildings that falls within each possible combination of labels (e.g. *good* and *bad*) is calculated. Then, the root mean square error between the particular method in a particular sample compared to the real values over all variable sets (vs) and all label sets (ls) is calculated. In other words, the correlation error using method m and sample i ($CE_{m,i}$) is calculated following Equation *(18)*:

$$CE_{m,i} = \sqrt{\frac{\sum_{vs=1}^{6}\sum_{ls=1}^{9}\left(p_{ls,m,i} - p_{ls,r}\right)^2}{54}} \tag{18}$$

where $p_{ls,m,i}$ is equal to the percentage of buildings that falls within this label set (e.g. *good* and *bad*) for this variable set (e.g. U-value roof and external wall) considering method m and sample i and $p_{ls,r}$ is equal to the percentage of buildings that falls within this label set for this variable set for the real values. The correlation error for method m is then defined as the median error over the 12 generated samples. The correlation error is actually a simplified version of the MMD where the 5-variate general distribution is transformed to pair-wise discrete distributions. While the correlation error is easier to understand, the MMD is much more complete.

To check the sensitivity of the correlation error, variations of this correlation error are assessed. First, instead of considering the U-values in sets of two, they can also be considered in sets of three or four, changing the number of variable sets from six to four or to one respectively. Second, instead of labelling the U-values in three bins (*bad*, *moderate* and *good*), they can also be labelled in five or ten bins, changing the number of label sets from 9 to 25 or to 100 respectively if the U-values are considered in sets of two. Third, instead of assessing the median error over the 12 samples, the minimum or the maximum error can also be assessed.

The number of random samples – in this work 12 – has been evaluated based on the variability of the correlation error over the different samples for the well-performing methods. The variability is found to be sufficiently small. Different numbers of samples have not been assessed.

The second additional performance indicator looks at the average behaviour of the different samples over all the buildings of the test dataset and is therefore referred to as the *average error*. Opposed to the two previous performance indicators, this indicator does not quantify the correlations between the variables within the generated samples, but rather the average behaviour of the generated samples. Methods that include the correlations in a good way often generate samples that are not as close to the average. They tend to generate more extreme and realistic scenarios, resulting in an increased average error. Therefore, this performance indicator is not equally important as the two previous performance

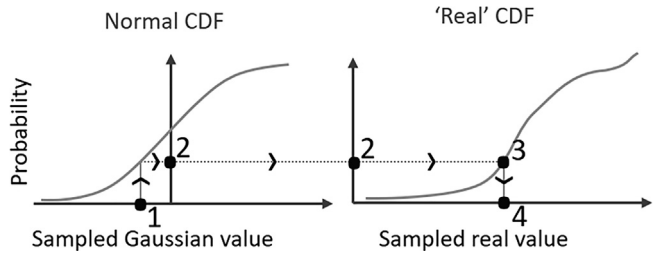

Fig. 5. Simplified illustration of how to obtain correlated samples using the Gaussian copula method. The illustration only shows one dimension of the multivariate distribution and should thus be applied for each of the variables of a dwelling. First, a correlated sample is drawn from a multivariate normal distribution. Then, these values are transformed to their corresponding percentiles. These percentiles are used to derive the real values based on the marginal distributions.

indicators. The average error of variable X using method m and sample i ($AE_{X,m,i}$) is calculated following Equation (19):

$$AE_{X,m,i} = \frac{|\bar{x}_{m,i} - \bar{x}_r|}{\bar{x}_r} \quad (19)$$

where $\bar{x}_{m,i}$ is equal to the average value over all the buildings of the test dataset of variable X using method m and sample i and $\bar{x}_r$ is equal to the real average value over all the buildings of the test dataset of variable X. The average error of variable X for method m is then defined as the median error over the 12 generated samples. The average error is calculated for both the UA-value on building level and the average U-value on building level (i.e. the UA-value divided by the total heat loss area), resulting in $AE_{UA,m}$ and $AE_{U,m}$ respectively.

To check the sensitivity of the average error, variations of the average error are assessed. Instead of assessing the median error over the 12 samples, the minimum or the maximum error can also be assessed.

Finally, next to these numerical performance indicators, the implementation effort and the ease of use after implementation of the different methods should be considered, which are more qualitative performance indicators.

3. Results and discussion

In this work, a probabilistic building characterisation method is proposed to allocate the U-values of the floor, external walls, windows and roof as well as the WWR to all single-family dwellings within UBEMs. First, QR is presented to generate the marginal distributions for these five variables. Then, different methods to build multivariate distributions from marginal distributions and to draw correlated samples are proposed. In this Section, these methods are compared extensively based on the proposed performance indicators. First, the performance of the marginal distributions is assessed. Then, the different methods to build multivariate distributions and to draw correlated samples are compared.

3.1. Accuracy of the marginal distributions

First, to obtain the marginal distributions of the five output variables, one QR model is fitted for each of the variables. The postal code, building type, construction year, total floor area, protected volume, ground floor area, façade area, and roof area are used to predict the distribution of the U-values of the floor, external walls, windows and roof as well as the WWR.

To illustrate the method, it is applied to a building of the test dataset of which the real values are known. The building is a terraced dwelling situated in Ypres (Belgium) and is constructed in 1962. Its floor area is 166 m^2, its protected volume is 491 m^3, its ground floor area is 96 m^2, its façade area is 195 m^2 and its roof area is 107 m^2. Fig. 6 displays the marginal distributions for the U-values and the WWR (in grey) as predicted by the QR models. The black lines in Fig. 6 represent the real values. As shown, the real values fall within the predicted distributions, illustrating an appropriate implementation of the QR models. Additionally, it should be noted that the predicted distributions inherently include the probability of particular renovations that might have taken place, which is an additional benefit of the QR models.

To check the accuracy of the QR models for the whole test dataset, rather than only for one particular building, the empirical coverage of the predictions at different prediction intervals is evaluated and shown in Table 2. The empirical coverage is close to the theoretical range for all output variables and all considered prediction intervals, as it deviates 0.6% at most. Therefore, the accuracy of the marginal distributions is concluded to be good.

3.2. Accuracy of the correlated samples

Subsequently, the different methods to build multivariate distributions from marginal distributions and to draw correlated samples are compared based on the four performance indicators that were introduced in Section 2.3. First, the results for the different variants of the sequential method are presented. Then, the results for the empirical copula method are discussed. Subsequently, the results for the Gaussian copula method are described. Finally, the best variants of the three methods are compared, while using the uncorrelated method as a reference.

3.2.1. Sequential method

First, the performance of the sequential methods is compared and shown in Fig. 7. Evaluation through MMD and CE both rank SM 2 and SM 1 at the top, although all variants of the sequential method perform rather similar compared to the other methods, which will be shown in Section 3.2.4. The MMD varies from 0.0068 to 0.0084 (Table 3). The CE, defined as the median over the 12 samples, varies from 0.0305 to 0.0333 (Table 3), i.e. the RMSE over all label sets (e.g. *good* and *bad*) and all variable sets (e.g. U-value roof and external wall) is 3.3% at most for the seven considered methods. The AE_U does not result in the same ranking as the MMD and the CE, nor as the AE_{UA}. However, again, there are no significant differences between the different methods according to the AEs. The AE_U and the AE_{UA}, both defined as the median over the 12 samples, vary from 0.0035 to 0.0049 and from 0.0032 to 0.0049 respectively (Table 3). The error of the average behaviour of the sequential methods is thus limited to 0.05. Both AEs result in slightly different rankings.

Additionally, the variability between the 12 samples is checked based on the CE and the AEs. In Fig. 7, the error bars represent the range between the minimal value and the maximal value of the 12 samples for both the CE and the AEs. The variability of the CE is very low, i.e. all samples respect the correlations of the real dataset to a similar extent. The variability of the AEs is higher, which can be expected as the 12 samples each represent a slightly different version of the considered building stock. Therefore, it is argued that the AEs are not the optimal indicators to assess the performance of a particular method to build multivariate distributions and to draw correlated samples. Nevertheless, the AEs provide an easy-to-understand average error for all methods, allowing to assess the overall behaviour of the different methods.

In Section 3.2.4, SM 2 and SM 1 are further compared to the other methods.

3.2.2. Empirical copula method

Second, the performance of the empirical copula methods is compared. In Fig. 8, the performance indicators are shown for the

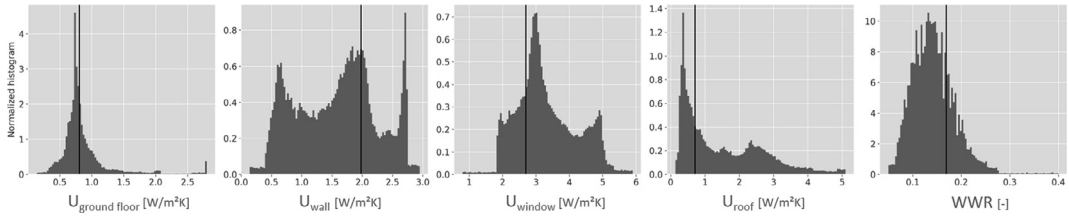

Fig. 6. Probability distributions predicted by the QR models for all the U-values and the WWR for one specific dwelling of the test dataset are indicated in grey. The real values for this dwelling are indicated by the black line. The terraced dwelling is situated in Ieper and is constructed in 1962. Its floor area is 166 m^2, its protected volume is 491 m^3, its ground floor area is 96 m^2, its façade area is 195 m^2 and its roof area is 107 m^2. To enhance readability, histograms are created based on 100,000 random samples from the CDF.

Table 2
Empirical coverages [%] on the 50%, 80%, 90% and 98% prediction interval for all output variables.

Output variable ↓ Prediction interval →	50%	80%	90%	98%
Ground floor U-value	50.3	80.2	90.0	98.1
External wall U-value	50.0	80.3	90.6	98.1
Window U-value	50.0	80.1	90.1	98.0
Roof U-value	50.2	80.0	90.1	98.0
WWR	49.7	80.3	90.1	98.1

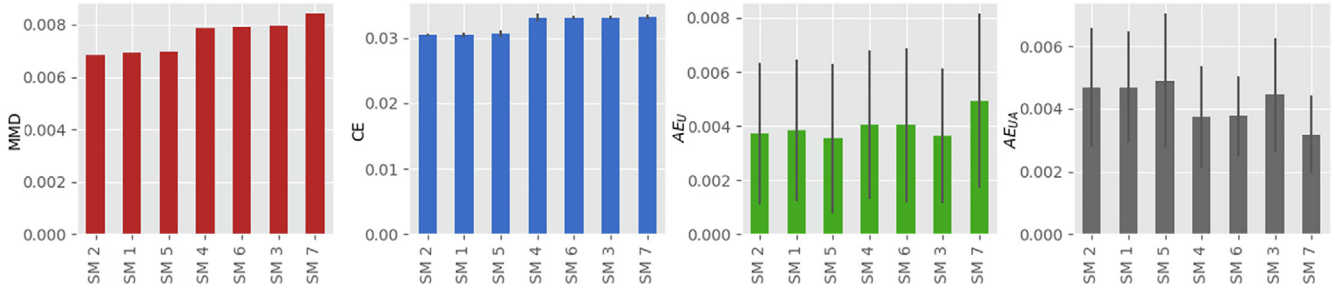

Fig. 7. Graphical overview of the four performance indicators for 7 of the 120 variants of the sequential method. The error bars show the range between the minimal and maximal values of the CE and the AEs over the 12 generated samples.

Table 3
Overview of the four performance indicators for the different methods to build multivariate distributions from marginal distributions and to draw correlated samples, ranked following the MMD.

	MMD	CE	AE_U	AE_{UA}
ECM Random	0.0020	0.0160	0.0044	0.0310
ECM Type	0.0024	0.0160	0.0050	0.0311
GCM	0.0027	0.0232	0.0026	0.0099
ECM Single	0.0036	0.0501	0.0781	0.0793
SM 2	0.0068	0.0305	0.0037	0.0047
SM 1	0.0069	0.0305	0.0038	0.0047
SM 5	0.0070	0.0307	0.0035	0.0049
ECM Year	0.0078	0.0242	0.0121	0.0234
ECM Type Year	0.0079	0.0240	0.0125	0.0233
SM 4	0.0079	0.0331	0.0041	0.0038
SM 6	0.0079	0.0331	0.0040	0.0038
SM 3	0.0079	0.0331	0.0036	0.0045
SM 7	0.0084	0.0333	0.0049	0.0032
Uncorrelated	0.0228	0.0415	0.0031	0.0041

considered variants of the empirical copula method, i.e. ECM Single, ECM Random, ECM Type, ECM Year and ECM Type Year. The MMD varies between 0.0020 and 0.0079 (Table 3). The CE varies between 0.0160 and 0.0501 (Table 3). Again, the MMD and the CE provide a similar ranking, except for ECM Single. The MMD does not capture the unreliability of ECM Single, as it does not distinguish between the 12 samples. The CE, however, does capture the unreliability of ECM Single, since it assesses the 12 samples separately and reports the median value. The reason for the unreliability of ECM Single will be explained further. According to the MMD, ECM Random and ECM Type perform three times better than ECM Year and ECM Type Year. The hypothesis is formulated that the subgroups according to the year and to the combination of type and year are too small and result in a distorted view. Additionally, subgroups according to year should be defined differently, as some historical events, such as the oil shocks of the 1970 s, caused a radical change in the construction industry. More in particular, the template Z for a building of 1975 contains buildings of 1970 until 1980. In their original state, some of these buildings will have some insulation, while others have no insulation at all. By now, some of these building will also have insulated during renovations. This hypothesis was checked visually. In the EPC dataset, the average U-value for the dwellings dropped around 1970 and then decreased steadily until 2007, although another drop is visible around 1985. However, these observations do not necessarily confirm the hypothesis. The AE_U provides a similar perspective than the CE, as opposed to the AE_{UA}. The AE_U and the AE_{UA} vary from 0.0044 to 0.0781 and from 0.0233 to 0.0793 respectively (Table 3). As mentioned earlier, the AEs should not be used to rank the different methods, but rather to check if the average behaviour of the generated samples is in accordance with the reference.

Additionally, the variability between the 12 samples is checked based on the CE and the AEs. In Fig. 8, the error bars show the range between the minimal value and the maximal value of the 12 samples for both the CE and the AEs. The variability of the CE is very low for all empirical copula methods, except for ECM Single. This observation makes ECM Single a rather unreliable method, as the

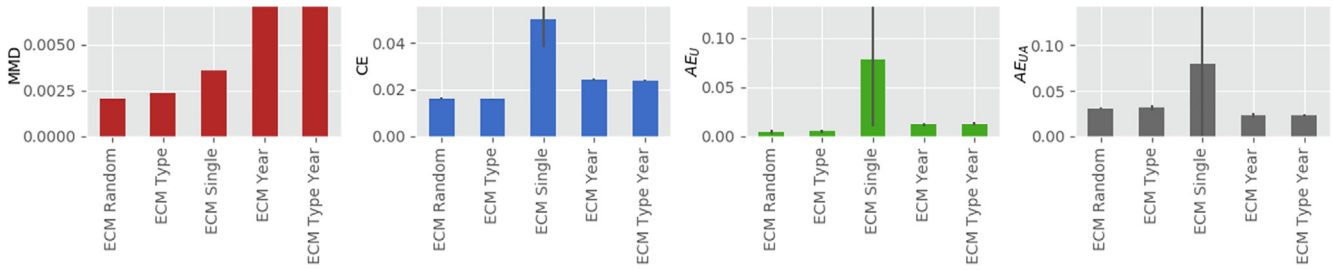

Fig. 8. Graphical overview of the four performance indicators for the empirical copula method. The error bars show the range between the minimal and maximal values of the CE and the AEs over the 12 generated samples.

correlations of the real dataset are not only not well respected but also highly depending on the particular sample. This can be expected as ECM Single only uses a single template Z to define the correlation of all buildings. Therefore, each sample highly depends on the selected buildings for template Z. The variability of the AE appears to be low for all empirical copula methods, except for ECM Single. However, this observation is only valid relative to ECM Single. Actually, the variability of the AE for the empirical copula methods is similar to the sequential methods, as will become more clear in Section 3.2.4, where ECM Random and ECM Type are compared to the other methods.

Although five variants of the empirical copula method have been proposed in this work, the definition of the template Z could still be improved, especially if the template Z is defined by drawing samples from a subset of similar buildings, as long as the subset remains sufficiently large. For example, the subset of similar buildings could be defined based on the k-nearest algorithm. However, this extension is left for future work.

3.2.3. Gaussian copula method

Third, the performance of the Gaussian copula method is discussed. The MMD and the CE of the Gaussian copula method are 0.0027 and 0.0232 respectively (Table 3). The AE_U and the AE_{UA} are 0.0026 and 0.0099 respectively (Table 3). The different performance indicators will be put into perspective in the next Section, by comparing to the other methods.

Opposed to the sequential methods and the empirical copula methods, no variations of the Gaussian copula method are considered, since this method already achieves good results compared to the other methods. The training dataset could have been subdivided into different groups (e.g. according to building type, construction year or both) and a covariance matrix could have been defined per group. Additionally, it could be explored how the covariance matrix varies from city centres to more rural context, but this would require more specific data.

3.2.4. All methods

Finally, the best performing variants of the different methods (i.e. ECM Random, ECM Type, GCM, SM 1 and SM 2) are compared in Fig. 9 and Table 3 to identify the most appropriate method to build multivariate distributions and draw correlated samples, which will be included in the probabilistic building characterisation method. To illustrate the added value of these methods, the uncorrelated method is included as well. According to the MMD, ECM Random, ECM Type and GCM perform very similar and are listed at the top. They perform eight to ten times better than the uncorrelated case. SM 1 and SM 2 are in between both. This is also reflected in the CE. While the sequential method is easy and simple to implement, it might not be able to characterise the correlation of the variables that are estimated first in a correct manner. In particular, while the QR model of the last estimated variable is conditioned on all the other random variables, the first estimated variable is only conditioned on the original inputs but on none of the other random variables. The different variants of the CE (i.e. considering the U-values in sets of three or four instead of two and labelling the U-values in five or ten bins instead of three bins) show similar results and are therefore not shown in this work.

The AEs, on the other hand, are larger for the empirical copula methods than for the sequential methods. Moreover, the AEs are very low for the uncorrelated case. This can be expected: samples that do not include correlation are more likely to generate values that are on average closer to the mean. Samples that include correlation are more likely to generate a more extreme scenario. However, as already mentioned, the average errors are only included to assess the average behaviour of the samples, as they do not include any form of correlation. The average errors are still sufficiently low for the best performing methods according to the MMD.

To identify the most appropriate method, two additional issues should be considered. First, the empirical copula methods are significantly easier to implement than the Gaussian copula method due to the simplicity to define the template Z. Second, although harder to implement, the Gaussian copula method is easier in use, as it requires less memory stored to generate multivariate distributions for new buildings. The Gaussian copula method only needs the covariance matrix that has been determined during setup, whereas the empirical copulas methods need the whole training dataset to sample historical buildings every time a new multivariate distribution is generated. Based on the four performance indicators and con-

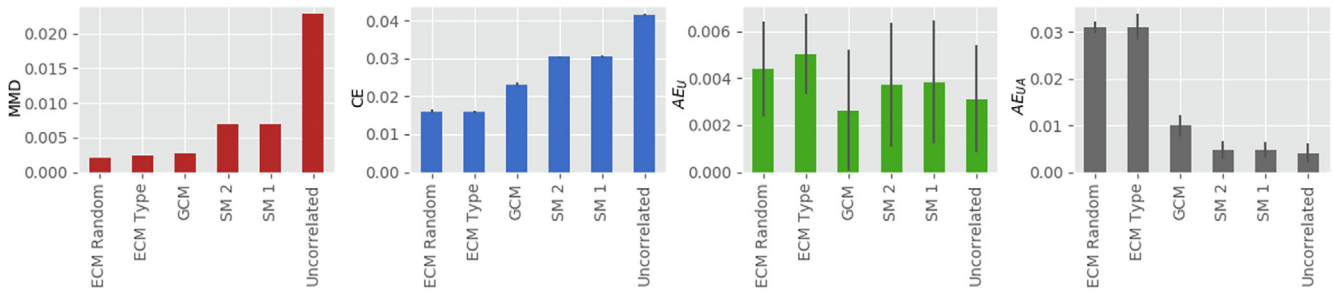

Fig. 9. Graphical overview of the four performance indicators for the best performing variants of the considered methods and the uncorrelated case. The error bars show the range between the minimal and maximal values of the CE and the AEs, over the 12 generated samples.

sidering the trade-off between implementation complexity and ease for future use, the GCM is put forward in this work as the most appropriate method to build multivariate distributions from marginal distributions and to draw correlated samples.

The probabilistic method will be particularly interesting to estimate realistic input data distributions to perform uncertainty and sensitivity analyses for the district energy demand of existing districts. However, not all sensitivity analysis methods are able to include the correlation between different input parameters. Only random samples can be drawn from these multivariate distributions. As a result, regression-based sensitivity analysis methods based on Monte Carlo approaches using optimised sampling or any other sensitivity analysis method that requires a specific sampling scheme cannot be used.

4. Conclusion

Urban building energy models are emerging, as they can be used to quantify the operational building energy use on district or city level as well as to estimate the impact of possible future scenarios. Buildings are often modelled following building or energy performance archetypes that define the energy performance-related data for all buildings following the particular archetype. As a result, the natural variability of the existing building stock is underestimated, causing some of the uncertainty on the simulation outcome. Therefore, in this work, a probabilistic building characterization method is proposed to model the full variability of the existing building stock within UBEMs. The method is able to estimate realistic input data distributions to perform uncertainty and sensitivity analyses for the district energy demand of existing districts. In this work, five parameters are considered, i.e. the U-values of the floor, external walls, windows and roof as well as the WWR, and are estimated based on data that is known for all Flemish single-family dwellings. The method is developed based on data of the Flemish energy performance certificates database.

First, QR has been proposed to generate the marginal distributions for the five variables. The accuracy of the marginal distributions is checked through the empirical coverage and is found to be good. The empirical coverage is close to the theoretical range for all output variables and all considered prediction intervals, as it deviates 0.6% at most.

Then, a method to build multivariate distributions from marginal distributions and to draw correlated samples has been developed and compared to two methods from literature that have been implemented for the first time within the context of the built environment in this work. Also, different variants of these three methods were shown. The *sequential method* has been proposed in this work and changes the input variables that are used to fit the QR models. The *empirical copula method* and the *Gaussian copula method* have been described in literature of different fields and were adapted to be used within the field of UBEMs in this work. Both methods employ a particular copula, i.e. a function that link multivariate distributions to their univariate marginal distributions.

To compare the different variants of these three methods, 12 samples are generated for 85,155 buildings of an out-of-sample test data set. Four performance indicators are proposed: the mean-maximum discrepancy (MMD), the correlation error (CE), the average error on the mean U-value and on the UA-value of the building (AEs). While the MMD is the most complete metric, the CE and AEs are proposed to better understand the differences between the different methods and the specific meaning of low and high MMD values. To calculate the CE, all U-values are discretized (i.e. labelled as *good*, *moderate* and *bad*) and combined in pairs of two (e.g. U-value roof and external wall). The CE quantifies the RMSE over all label sets (e.g. *good* and *bad*) and all variable sets (e.g. U-value roof and external wall) in a particular sample and a particular method compared to the real values. The CE is then defined as the median over the 12 samples. Additionally, the AE quantifies the absolute percentage error on the average behaviour of all buildings in a particular sample and a particular method compared to the real values. The AE is defined as the median over the 12 samples.

According to the MMD, the Gaussian copula method and the empirical copula methods, where the copula is defined based on random buildings or on buildings of the same building type, are found to perform best. Their MMDs are eight to ten times lower that the MMD of the uncorrelated method. For these methods, the CE varies from 1.6% to 2.3% and the AE is limited to 3%. For the empirical copula method, it is important that the template Z is defined based on sufficient data points from the training dataset. This is the reason why the empirical copula methods, where the copula is defined based on 12 random buildings or on buildings of the same construction period or on buildings of the same construction period and the same type, do not perform as good. Additionally, the sequential method, developed in this work, does not perform as good as the Gaussian and the empirical copula method since not all correlations are fully included (i.e. the QR model for the last estimated variable is conditioned on all other variables, but the first does not include any correlation).

Based on the four numerical performance indicators and considering the implementation complexity and the ease for future use, the Gaussian copula method is put forward as the preferred method to build multivariate distributions from marginal distributions and to draw correlated samples and is included in the probabilistic building characterisation method.

The probabilistic building characterisation method can be used to feed data into UBEMs for building-level parameters. The method is particularly interesting to obtain realistic input building-level parameter variations to perform uncertainty and sensitivity analyses of the energy demand for existing residential neighbourhoods within future work. The uncertainty analysis will show the uncertainty on the simulated district energy demand that currently is to be expected while using the best available input data without intensive on-site data collection. The sensitivity analysis will allow to identify the most influencing parameters for the district energy demand. Ideally, these parameters will be collected with more care to decrease the uncertainty within future district energy simulations. In addition to the uncertainty and sensitivity analysis, it will be investigated what the impact of uncertainty on the district energy demand is within different use cases: the design of district heating systems (e.g. sizing storage units), the operation of districts where heating is supplied by a district heating system or by heat pumps (e.g. control of the system) and the optimal renovation strategy to achieve energy positive district and cities. As UBEMs are used to answer important questions within the context of achieving energy neutral or positive districts and cities, it is crucial to assess the impact of uncertainty.

CRediT authorship contribution statement

Ina De Jaeger: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Jesus Lago:** Conceptualization, Methodology, Software, Validation, Writing - review & editing. **Dirk Saelens:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgements

The authors gratefully acknowledge the Research Foundation Flanders (FWO) and the Flemish Institute for Technology (VITO) for funding this research. Ina De Jaeger holds a PhD grant fundamental research financed by the Research Foundation - Flanders (FWO) and the Flemish Institute for Technological Research (VITO) (grant number: 11D0318N). This research has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 675318 (INCITE). Additionally, the authors are thankful to the Flemish Energy Agency (VEA) for providing the EPC data of Flanders.

References

[1] M. Le Guen, L. Mosca, A.T.D. Perera, S. Coccolo, N. Mohajeri, J.-L. Scartezzini, Improving the energy sustainability of a Swiss village through building renovation and renewable energy integration, Energy Build. 158 (2018) 906–923, https://doi.org/10.1016/j.enbuild.2017.10.057.

[2] K. Orehounig, G. Mavromatidis, R. Evins, V. Dorer, J. Carmeliet, Towards an energy sustainable community: An energy system analysis for a village in Switzerland, Energy Build. 84 (2014) 277–286, https://doi.org/10.1016/j.enbuild.2014.08.012.

[3] J. Hirvonen, J. Jokisalo, J. Heljo, R. Kosonen, Towards the EU emissions targets of 2050: optimal energy renovation measures of Finnish apartment buildings, Int. J. Sustain. Energy. 38 (2019) 649–672, https://doi.org/10.1080/14786451.2018.1559164.

[4] T.M. Lawrence, M.-C. Boudreau, L. Helsen, G. Henze, J. Mohammadpour, D. Noonan, D. Patteeuw, S. Pless, R.T. Watson, Ten questions concerning integrating smart buildings into the smart grid, Build. Environ. 108 (2016) 273–283, https://doi.org/10.1016/J.BUILDENV.2016.08.022.

[5] J. Kensby, A. Trüschel, J.-O. Dalenbäck, Potential of residential buildings as thermal energy storage in district heating systems – Results from a pilot test, Appl. Energy. 137 (2015) 773–781, https://doi.org/10.1016/J.APENERGY.2014.07.026.

[6] C.F. Reinhart, C. Cerezo Davila, Urban building energy modeling – A review of a nascent field, Build. Environ. 97 (2016) 196–202, https://doi.org/10.1016/j.buildenv.2015.12.001.

[7] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, Build. Environ. 45 (2010) 1683–1697, https://doi.org/10.1016/j.buildenv.2010.01.021.

[8] I. De Jaeger, G. Reynders, C. Callebaut, D. Saelens, A building clustering approach for urban energy simulations, Energy Build. 208 (2020) 109671, https://doi.org/10.1016/j.enbuild.2019.109671.

[9] C. Protopapadaki, D. Saelens, Heat pump and PV impact on residential low-voltage distribution grids as a function of building and district properties, Appl. Energy. 192 (2017) 268–281, https://doi.org/10.1016/j.apenergy.2016.11.103.

[10] D. Müller, A. Monti, S. Stinner, T. Schlösser, T. Schütz, P. Matthes, H. Wolisz, C. Molitor, H. Harb, R. Streblow, Demand side management for city districts, Build. Environ. 91 (2015) 283–293, https://doi.org/10.1016/j.buildenv.2015.03.026.

[11] G. Mavromatidis, K. Orehounig, L.A. Bollinger, M. Hohmann, J.F. Marquant, S. Miglani, B. Morvaj, P. Murray, C. Waibel, D. Wang, J. Carmeliet, Ten questions concerning modeling of distributed multi-energy systems, Build. Environ. 165 (2019), https://doi.org/10.1016/j.buildenv.2019.106372.

[12] B. Morvaj, R. Evins, J. Carmeliet, Optimising urban energy systems: Simultaneous system sizing, operation and district heating network layout, Energy. 116 (2016) 619–636, https://doi.org/10.1016/j.energy.2016.09.139.

[13] R. Wu, G. Mavromatidis, K. Orehounig, J. Carmeliet, Multiobjective optimisation of energy systems and building envelope retrofit in a residential community, Appl. Energy. 190 (2017) 634–649, https://doi.org/10.1016/j.apenergy.2016.12.161.

[14] K. Orehounig, G. Mavromatidis, R. Evins, V. Dorer, J. Carmeliet, Predicting Energy Consumption of a Neighborhood Using Building Performance Simulations, Build. Simul. Optim. Conf., 2014.

[15] I. De Jaeger, G. Reynders, D. Saelens, Impact of spatial accuracy on district energy simulations, Energy Procedia. 132 (2017) 561–566, https://doi.org/10.1016/j.egypro.2017.09.741.

[16] D. Cuypers, B. Vandevelde, M. Van Holm, S. Verbeke, Belgische woningtypologie: nationale brochure over de TABULA woningtypologie, 2014. http://episcope.eu/fileadmin/tabula/public/docs/brochure/BE_TABULA_TypologyBrochure_VITO.pdf (accessed April 11, 2018).

[17] G. Mavromatidis, K. Orehounig, J. Carmeliet, A review of uncertainty characterisation approaches for the optimal design of distributed energy systems, Renew. Sustain. Energy Rev. 88 (2018) 258–277.

[18] N. Ghiassi, A. Mahdavi, Reductive bottom-up urban energy computing supported by multivariate cluster analysis, Energy Build. 144 (2017) 372–386, https://doi.org/10.1016/j.enbuild.2017.03.004.

[19] G. Tardioli, R. Kerrigan, M. Oates, J. O'Donnell, D.P. Finn, Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach, Build. Environ. 140 (2018) 90–106, https://doi.org/10.1016/J.BUILDENV.2018.05.035.

[20] J. Sokol, C. Cerezo Davila, C.F. Reinhart, Validation of a Bayesian-based method for defining residential archetypes in urban building energy models, Energy Build. 134 (2017) 11–24, https://doi.org/10.1016/j.enbuild.2016.10.050.

[21] T. Hong, Y. Chen, X. Luo, N. Luo, S.H. Lee, Ten questions on urban building energy modeling, Build. Environ. 168 (2020), https://doi.org/10.1016/j.buildenv.2019.106508.

[22] C.S. Monteiro, C. Costa, A. Pina, M.Y. Santos, P. Ferrão, An urban building database (UBD) supporting a smart city information system, Energy Build. 158 (2018) 244–260, https://doi.org/10.1016/j.enbuild.2017.10.009.

[23] M. Österbring, É. Mata, L. Thuvander, M. Mangold, F. Johnsson, H. Wallbaum, A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model, Energy Build. 120 (2016) 78–84, https://doi.org/10.1016/j.enbuild.2016.03.060.

[24] A. Mastrucci, P. Pérez-López, E. Benetto, U. Leopold, I. Blanc, Global sensitivity analysis as a support for the generation of simplified building stock energy models, Energy Build. 149 (2017) 368–383, https://doi.org/10.1016/j.enbuild.2017.05.022.

[25] J. Keirstead, M. Jennings, A. Sivakumar, A review of urban energy system models: Approaches, challenges and opportunities, Renew. Sustain. Energy Rev. 16 (2012) 3847–3866, https://doi.org/10.1016/j.rser.2012.02.047.

[26] M. Kavgic, D. Mumovic, A. Summerfield, Z. Stevanovic, O. Ecim-Djuric, Uncertainty and modeling energy consumption: Sensitivity analysis for a city-scale domestic energy model, Energy Build. 60 (2013) 1–11, https://doi.org/10.1016/j.enbuild.2013.01.005.

[27] A.T.D. Perera, S. Coccolo, J.L. Scartezzini, D. Mauree, Quantifying the impact of urban climate by extending the boundaries of urban energy system modeling, Appl. Energy. 222 (2018) 847–860, https://doi.org/10.1016/j.apenergy.2018.04.004.

[28] F. Biljecki, H. Ledoux, J. Stoter, An improved LOD specification for 3D building models, Comput. Environ. Urban Syst. 59 (2016) 25–37, https://doi.org/10.1016/j.compenvurbsys.2016.04.005.

[29] R. Koenker, G. Bassett, Regression Quantiles, Econometrica. 46 (1978) 33, https://doi.org/10.2307/1913643.

[30] M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The Schaake Shuffle: A Method for Reconstructing Space-Time Variability in Forecasted Precipitation and Temperature Fields, J. Hydrometeorol. 5 (2004) 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

[31] D.S. Wilks, Multivariate ensemble Model Output Statistics using empirical copulas, Q. J. R. Meteorol. Soc. 141 (2015) 945–952, https://doi.org/10.1002/qj.2414.

[32] P. Pinson, H. Madsen, H.A. Nielsen, G. Papaefthymiou, B. Klöckl, From probabilistic forecasts to statistical scenarios of short-term wind power production, Wind Energy. 12 (2009) 51–62, https://doi.org/10.1002/we.284.

[33] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A Kernel Two-Sample Test, J. Mach. Learn. Res. 13 (2012) 723–773.