Data-driven building archetypes for urban building energy modelling

Oleksii Pasichnyi a, *, Jörgen Wallin b, Olga Kordas a

a Urban Analytics and Transitions Research Group, Department of Sustainable Development, Environmental Sciences and Engineering, School of Architecture and the Built Environment, KTH Royal Institute of Technology, Sweden
b Urban Analytics and Transitions Research Group, Department of Energy Technology, School of Industrial Engineering and Management, KTH Royal Institute of Technology, Sweden

A R T I C L E   I N F O

Keywords:
Building archetype
Urban building energy modelling
Building retrofitting
Electric heating
Stockholm

A B S T R A C T

This paper presents an approach for using rich datasets to develop different building archetypes depending on the urban energy challenges addressed. Two case studies (building retrofitting and electric heating) were analysed using the same city, Stockholm (Sweden), and the same input data, energy performance certificates and heat energy use metering data. The distinctive character of these problems resulted in different modelling workflows and archetypes being developed. The building retrofitting case followed a hybrid approach, integrating statistical and physical perspectives, estimating energy savings for 5532 buildings from seven retrofitting packages. The electric heating case provided an explicitly statistical data-driven view of the problem, estimating potential for improvement of power capacity of the local electric grid at peak electric power of 147 MW. The conclusion was that the growing availability of linked building energy data requires a shift in the urban building energy modelling (UBEM) paradigm from single-logic models to on-request multiple-purpose data intelligence services.

© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Energy conservation in the existing building stock is the cornerstone for reducing primary energy consumption and greenhouse gas (GHG) emissions in cities [1]. In order to meet ambitious energy efficiency and climate mitigation targets, cities need to understand current energy demand and future effects of various retrofitting decisions. At the same time, the electrification of heating is recognised as one of the main technological changes driving reduction of GHG emissions within the heating sector [2]. However, the increasing grid loads associated with space heating impose higher requirements on feeder and transformer capacity, affecting the power quality, capacity, cost and regulation requirements in the distribution networks [3].

Various energy models are used for planning and decision support for strategies on building stock retrofitting and securing electric power supply. However, the conventional energy models are either limited in detail or too exhaustive to be applied on a large scale [4]. At the same time, emerging urban building energy models (UBEMs) use growing volumes of urban energy data and significantly reduce the amount of effort needed from human modellers [5]. UBEMs provide automated generation of building energy models through abstraction of building stock by different ‘building archetypes’, i.e. sample or virtual buildings that characterise subsets of buildings of the same kind [6]. However, replacement of the whole archetype building subset by a single building induces a risk of oversimplification, resulting in poor quality of the modelling output. Availability of rich datasets on factual heat energy use opens the way for approaches to generating various building archetype subsets based on physical and statistical models of the city building stock [7].

The aim of this study was to demonstrate how using rich datasets allows different building archetypes to be developed depending on the urban energy challenges addressed. It builds on two cases from the same city, Stockholm (Sweden), using the same input data, energy performance certificates and heat energy use metering data. The purpose of Case 1 was to identify the potential for energy saving through large-scale building energy retrofitting. The purpose of Case 2 was to explore the current usage of electric heating and identify its potential for reducing the electricity power demand. A UBEM for Stockholm with heating as the main scope was developed in both cases, but the distinctive character of the problems addressed resulted in different modelling workflows and archetypes being developed. In Case 1, future effects of retrofitting...
were estimated, while Case 2 provided statistical aggregation for building electricity use for heating that exists now. Hence, Case 1 followed a hybrid approach, combining statistical (energy signatures) and physical (building energy simulations) modelling, while Case 2 corresponded to a statistical view of the problem built explicitly on currently available data. Regarding the archetypes developed, Case 1 did not have a requirement to cover the whole city. Instead, the focus was narrowed down to three archetypes corresponding to the most typical classes of buildings, ensuring feasibility of the retrofitting measures analysed. In Case 2, all buildings using electric heating were the focus and the aim of archetyping was to provide several levels of detail, allowing for comprehensive exploration of the power capacity issue.

The study examined whether integration of various building energy datasets, such as energy performance certificates and metered heat energy use, allows for greater flexibility in the choice of UBEM, resulting in better alignment of the archetypes constructed to the problem addressed and consequently better decisions.

2. Background

2.1. Urban building energy modelling

Rapid development of computing and sensor technologies has created a ‘big data’ challenge to running building energy simulations in a conventional way, which calls for new approaches for handling and utilising the data [8]. A comprehensive overview of types of existing building energy data and methods for their collection are described by Mantha et al. [9]. UBEM has emerged in recent years as an efficient hybrid of top-down statistical and bottom-up engineering approaches [4,10]. It is expected to become a main planning tool for energy utilities, municipalities, urban planners and other professionals [5].

Two main UBEM approaches can be identified from the literature: physical modelling [5] and data-driven modelling [11,12] compared these two major UBEM approaches through analysis of performance for physical GIS-based UBEM with detailed dynamic building energy simulation in IDA ICE software and data-driven UBEM based on nonlinear energy signatures (ES). They concluded that both approaches have advantages and disadvantages. Therefore, choice of UBEM approach should be guided by the requirements for the modelling outcomes and available data and resources for each particular case.

Creation of city-wide models requires the synthesis of many models with similar characteristics, which is not possible to do manually. Therefore, UBEMs utilise building archetypes, an approach that provides a compromise between accuracy and speed of simulation [13]. It has been extensively used in the context of national and regional bottom-up building stock models to analyse the current state of building stock and the aggregated impact of energy efficiency policies or new technologies [14–16]. Archetyping is usually performed in two stages: segmentation (also referred as classification) and characterisation (also referred as parameterisation) [5]. Beside segmentation and characterisation performed for archetyping, data preparation, modelling and aggregation (quantification) stages are usually present in UBEMs.

Segmentation is a process of introducing taxonomy to the population of buildings analysed, classifying them into one or several layers of categories. This allows archetypes to be defined either as classes (subsets) of similar buildings, or as typical (sample or virtual) representatives of these classes (archetype buildings). The conventional approach to segmentation is to split the building stock according to schemes defined by modeller in a manual (deterministic) or semi-automatic (statistic) way, usually sticking to splits by socio-economic (type of building usage, income level) [17,18], spatial (climate zone, location) [19], structural (e.g. age, floor area, envelope form, number of floors) [20], energy installation (e.g. heating source, ventilation system, status of refurbishment) [21] or performance (e.g. energy use intensity, total energy, peak power) [22,23] features. However, the development of machine learning techniques (both supervised and unsupervised) allows for more automated statistic segmentation, where the modeller’s main input is to define the building (energy) similarity metrics to be used, and not the feature splits themselves. For instance, application of clustering techniques for archetyping has been demonstrated [24,25].

Characterisation is a process of developing representation for each archetype, defining values for the relevant parameters in a deterministic (single value) or probabilistic (distribution) way (see Loga et al. [26] and Burke et al. Burke et al. (2017)). The scope of parameter characterisation depends on the data available, problem addressed and type of building energy performance model used. The values for characterisation are obtained from building data, literature and building surveys [27]. Absent data or data of insufficient granularity can lead to oversimplified and biased archetype characterisation, in which case various calibration methods should be applied [28,29].

Archetyping approaches have received much attention recently, mainly due to their use as a core of UBEMs. However, their purpose and functions can be seen as wider than a bare simplification of the building stock for the purpose of further modelling. The following
functions of archetypes can be identified in the literature:

a. Aggregation – provide an overview of the urban building stock analysed [6,23].

b. Explanation – highlight the triggers for the energy performance of the urban building stock through interpretative taxonomies [21].

c. Estimation – estimate energy performance of the urban building stock with the quality of the estimate being within the accepted limits [22].

d. Simplification – decrease computational complexity through removing data redundancy [22].

e. Framing – address only those segments of the building stock that are relevant for the problem addressed [30].

f. Compression – maximise the amount of preserved information about the building stock [22].

g. Transferability – apply archetypes from other cases if input data are lacking [28].

h. Benchmarking – compare building energy performance and related strategies for different cases [31].

Case studies were performed on two archetyping processes that both had Stockholm as the target city, with the same input data used. The next section provides the context and describes the focus of the cases analysed.

2.2. Stockholm case studies

Stockholm metropolitan area is one of the fastest growing urban areas in Europe [32]. The City of Stockholm aims to become a fossil-free city by 2040 [33] which is an even higher ambition than the overall Swedish national climate target of becoming a fossil-free state by 2050 [34]. Stockholm has an extensive district heating (DH) system that covers around 85% of the total heat energy demand in the city. The system is one of the most advanced multi-energy systems in the world and largely corresponds to “fourth-generation DH” [35]. The largest actor in the system is AB Stockholm Exergi (co-owned by Fortum Group), producing around 8 TWh annually [36], and operating 15 068 heat supply metering points. The electrification of heating highlighted as a trend by Ref. [2] is also evident in Stockholm, greatly increasing the role of heating in power grid capacity.

Both cases analysed in this paper were limited to the heat provision in Stockholm building stock. Case 1 explored the potential for reducing heat energy demand through building retrofitting. Case 2 explored electricity-based heating and its potential to increase the capacity of the local power grid.

2.2.1. Case 1. building retrofitting

Heating and cooling of the building stock is associated with approximately 40% of total GHG emissions in Stockholm. Hence a fundamental target for the City of Stockholm is to decrease building energy use. Stricter requirements on new buildings is part of the strategy. However, since most of the city’s building stock already exists, in order to achieve its environmental goals the city intends to introduce energy efficiency measures (varying for different categories of buildings) that will achieve an average decrease of 30% in energy use for heating and cooling [37].

The purpose of Case 1 was to estimate the energy savings potential of large-scale retrofitting, demonstrated for several typical groups of buildings in Stockholm.

2.2.2. Case 2. electric heating

The projected population increase for Stockholm from 800 000 inhabitants in 2008 to 1 200 000 in 2040 imposes strong requirements on development of urban infrastructure [38]. Large investments are being made for rebuilding of the electricity grid in Stockholm region, including construction of new substations and strengthening of existing power lines. This would improve the total grid capacity and allow all city needs to be met, as depicted in Fig. 1. However, the expansion is expected to be accomplished not earlier than 2027, resulting in a noticeable gap between currently available grid capacity and projected electricity demand within the period 2021–2026. If this deficit in grid capacity were to come into effect, it would not only stress regular city operations, imposing hard competition for secure power supply over acting services, but also jeopardise any further city developments dependent on increased power supply. For instance, it would put under risk construction of new urban districts, transition of the transport fleet to electrical vehicles, deployment of more data centres, emergence of new advanced industries relying on electricity, and, consequently, the whole ambition of Stockholm for fast transition to a fossil-free city.

This potential power capacity deficit could be mitigated through either increasing in-city electricity production (e.g. installation of photovoltaics plus storage systems or more power production at combined heat and power plants (CHPs)) or reducing electricity consumption by energy efficiency measures (e.g. switch to LED lighting, load shift through local energy storage and demand response mechanisms) or replacement of electricity by another energy source (e.g. use of DH for indoor heating & cooling). As Stockholm City has a well-developed DH system, switching electricity-based heating buildings to DH can be a possible solution to reduce total electrical power load demand.

The purpose of Case 2 was to explore the building stock using electricity for heating and estimate the potential for decreasing electricity power demand by changing the source of heat for Stockholm’s electricity-heated building stock to another energy carrier.

3. Methodology

The general research design followed the case study method [40]. Two cases were studied: Case 1 addressed the potential for improving building energy efficiency through large-scale building retrofitting, while Case 2 addressed the potential for freeing up power supply capacity through reducing electric heating. While both cases share most of their context (the same city (Stockholm), the same general domain (urban energy) and the same input data), the problems addressed were quite different and can have implications for distinct areas of urban planning and decision making. This resulted in a marked difference in the branches of the urban building modelling workflow for the two cases (Fig. 2).

In this study, parts of UBEM workflow developed in a previous study for analysing strategies for large-scale retrofitting [41] were used. Case 1 followed a hybrid approach, integrating both statistical (model) and physical (simulation) perspectives on the building stock analysed, while Case 2 provided an explicitly statistical view on the problem, and hence insights obtained on the buildings analysed were purely data-driven. Thus the input data (Section 3.1) and statistical modelling (Section 3.2) were the same, but segmentation (Section 3.4), characterisation (Section 3.5) and aggregation (Section 3.7) were conducted differently in each case.

---

1 Stockholm municipality is referred to hereafter as “Stockholm”.
2 The total area of currently existing building stock in Stockholm is 64 million m².

---

https://www.stockholmsstrom.net.
Furthermore, additional data transformation (Section 3.3) and building energy simulations (Section 3.5) were performed in only one of the cases. All calculations were performed in R language [42] in the RStudio IDE [43], with a number of additional packages applied: gridExtra, Metrics, tidyverse [44], UpSetR [45] and VennDiagram. Building energy simulations were conducted in the EnergyPlus [46] energy simulation engine through an interactive interface DesignBuilder [47].

3.1. Input data

Two main data sources served as input data for both cases:

1. Energy performance certificates (EPC) data were used for building stock segmentation and further accounting of the heat consumption. The EPC dataset was obtained from the Board of Housing, Building and Planning (Boverket). This EPC dataset covered 30,472 buildings in Stockholm and contained information on building type, construction year, floor area, envelope form, energy use per source, energy system installations etc. EPCs are mandatory for all buildings that are: a) subject to sale or rent; b) frequently visited by the public; or c) newly constructed.

2. Measured factual heating data (FDH) on energy consumption (hereafter measured or FDH) were used for creating statistical models for each building analysed. The measured data for the case were extracted from the district heating hourly metering dataset obtained from the local energy utility AB Stockholm Exergi. The FDH data covered the year 2012 with hourly precision, representing all customers of the local district heating network. The data also contained information on type of building usage, building age category and total heated area $A_{\text{temp}}$.

The EPC and measured datasets were linked through matching cadastral codes, as described in Ref. [48]. Buildings that failed to link or were identified as having significant inconsistency (deviation of more than 20%) between EPC and FDH data were removed before the statistical modelling stage (Section 3.2), as it is heavily dependent on FDH data.

Beside EPC and measured data, several additional data sources were used:

- Climate data on the ambient temperature were used for fitting the statistical models of the building stock and estimating the heat power demand in both cases. Time series for 2012 and the period 1981–2010 with hourly precision were used [49].
- Reference data on standardised use and building envelope information were used to characterise additional features required for setting up virtual building energy simulation model archetypes in the retrofitting case. Here the reference data were
obtained from the Sveby project [50] and the “Så bygges husen” book [51].

3.2. Statistical models

Statistical modelling of heat demand was used to estimate building energy performance from measured data in both cases. Heating power demand of each building was modelled with energy signatures $ES = (\theta, c, \beta)$ obtained from fitting the specific heat load $q(t)$ calculated from the FDH data with the following quasilinear regression model:

$$\frac{Q(t)}{A_{\text{temp}}} = q(t) = \begin{cases} 
  c + \beta (\theta - t), & t < \theta \\
  c, & t \geq \theta 
\end{cases} \quad (1)$$

where $t$ is ambient temperature, $\theta$ is balance point temperature, $c$ is base load (domestic hot water consumption) and $\beta$ is energy performance coefficient. Fig. 3 exemplifies energy signature for one of the modelled buildings in Stockholm.

Energy signature was selected as a simple statistical model that requires only energy use data, and supports comparability across large numbers of dwellings [7]. The energy signature method and its applications have been addressed in a number of studies. (e.g. Refs. [52–59]. One of those studies [59] concluded that the method is robust, with deviation of less than 2% in the estimated heat loss factor (SPF) and its applications have been addressed in a number of studies. (e.g. Refs. [52–59]. One of those studies [59] concluded that the method is robust, with deviation of less than 2% in the estimated heat loss factor (SPF) and $R^2$ (coefficient of determination), as recommended by Ruiz and Bandera [60]. However, since the measured data contained a significant amount of meterings with zero values, MAPE was replaced by MASE (Mean Absolute Scaled Error), as proposed by Hyndman and Koehler [61].

3.3. Data transformation

To perform the analysis of electricity-based heat energy demand in Case 2 (Section 4.3), the values of supplied energy reported in EPC data were converted into used (heat) energy following assumptions about seasonal performance factors (SPF) for various heat sources used by the energy auditors during EPC data collection (Table 1):

$$Q_{sourcej} = \eta_{sourcej} \times E_{sourcej}$$

where $Q_{sourcej}$ is amount of heat, $\eta_{sourcej}$ is seasonal performance factor (SPF) and $E_{sourcej}$ is amount of purchased energy.

In the case of heat pump sources, the distinction was made for the SPF assumptions depending on the building type. High-temperature systems ($\eta_{GSHP} = 3.1, \eta_{ASHPW} = 2.6$) were assumed for all buildings except single-family dwellings, which were calculated with the assumption of 50/50 mix of high-temperature and low-temperature systems ($\eta_{GSHP} = 2.1 + 2.8 = 3.5, \eta_{ASHPW} = 2.6 + 2.2 = 2.85$). Finally, for EAHPs a new type ($\eta_{EAHP} = 2.6$) was selected, as this type of installation has greatly expanded recently and is expected to further dominate the market.

In addition, Boolean values for using/non-using a particular heat source were calculated for each building:

$$QB_{sourcej} = \begin{cases} 
  0, Q_{sourcej} = 0 \\
  1, Q_{sourcej} > 0 
\end{cases} \quad (3)$$

Boolean representation of used heat sources was required to perform segmentation with heat sources as one of the splitting features.

3.4. Segmentation

All segmentation schemes applied followed the logic of rule-based categorisation, where each building can belong exactly to one of a set of classes (archetypes):

$$\forall \text{building, } \exists! k : \text{building} \in A_k$$
spaces with all possible heat mixes, used in the scheme (Table 2).

where together or kept separately, deriving a number of grouping schemes
detail and complexity, the heat sources (Table 1) were grouped
schools, hospitals, etc.)). Further, various segmentations by types of
archetypes could be an empty set, \( \mathcal{A} = \emptyset \) (the whole population
is required to be covered with archetypes) or not (only a subset is
relevant for archetyping).

The archetypes \( \{A_i\} \) were defined through imposing mutually exclusive
sets of conditions on the building features in EPC data.

In Case 1 (building retrofitting), the archetyping process was
mainly driven by expert knowledge and therefore the resulting
archetypes were defined directly through conditioning by usage
type of buildings, their age and connectivity to DH network, leaving
the remaining part of the city beyond the scope of the study:

\[
A_k \subset A = A_1 \cup A_2 \cup \ldots \cup A_n, \quad A_i \cap A_j = \emptyset, \quad \forall i, j = 1 \ldots n, i \neq j
\] (4)

where \( A_k \) is arbitrary archetype and \( A = A_1 \cup A_2 \cup \ldots \cup A_n \).

Depending on the requirements for archetyping, the residual
archetype could be an empty set, \( \mathcal{A} = \emptyset \) (the whole population
is required to be covered with archetypes) or not (only a subset is
relevant for archetyping).

In Case 2 (electric heating), the archetyping was conducted in a
trial and error manner, updating the segmentation scheme until an
acceptable balance between the complexity and applicability was
found.

Splitting by three general building usage types \( B_3 = \{S,M,O\} \)
was initially applied, to represent the difference in energy performance
due to type of building use (\( S = \) single family dwellings,
\( M = \) multi-residential buildings, and \( O = \) other buildings
(officials, schools, hospitals, etc.)). Further, various segmentations by types of
heat sources were elaborated. Depending on the required level of
detail and complexity, the heat sources (Table 1) were grouped
together or kept separately, deriving a number of grouping schemes
\( H_n \) where \( n = \{3, 6, 8, 13\} \) is the number of types of heat sources
used in the scheme (Table 2).

These grouping schemes served as the basis for generation of
spaces with all possible heat mixes, defined as a power set over the
heat source grouping scheme:

\[
\mathcal{P}(H_n) = \{0,1\}^{H_n}
\] (6)

In this way, the maximum possible number of subsets (all possible combinations
of \( n \) heat sources from the scheme \( H_n \)) would be \( |\mathcal{P}(H_n)| = 2^n - 1 \) (\( \emptyset \) is skipped as not applicable in this case).

For the purpose of simplicity, hereafter we refer to any mix
element of such sets of mixes by enumerating only the types of heat
sources used \( (Q_{\text{source}} = 1) \). For example, for the grouping scheme
\( H_3 = \{DH,FBH,EH\} \), the number of possible mixes is \( |\mathcal{P}(H_3)| = 2^3 - 1 = 7 \) and the buildings using both DH and EH would belong to the
subset \( DH = 1, FBH = 0, EH = 1 \) or, if using short notation \( DH, EH \) or
\( DH+EH \), which is the element of \( \mathcal{P}(H_3) \).

Two main groups of heat mixes involving EH were distinguished
in \( \mathcal{P}(H_3) \) to perform the detailed analysis on the amount of heat
consumed and associated power demand per combination of building types (B3, Table 3)
and electric heat sources (H6, Table 2). The first group, “EH mainly” (\( G_{EH} = EH \cup \{DH\} \)) stands for the buildings using electricity, both solely and in mixture
with fuel burners (FBH), but without district heating (DH). Thus group \( G_{EH} \) unites all buildings that are potentially harder targets to be
switched from EH, as they are not connected to the DH system, and
FBH is often not a better alternative to consider due to the higher
local emissions. The second group, “EH partly” (\( G_{EH} = EH \cup \{DH\} \)), unites all buildings using both electricity and district heating, whether with or without
FBH. Hence, group \( G_{EH} \) unites all buildings that have good potential
for switching, as they are already connected to the DH network, so
no or small capital investments are associated with the switch and
it can be achieved in a short time through proper economic
incentives.

Finally, the resulting archetypes were defined though segmentation
of buildings by Cartesian product of three usage types B3 and
two target groups of heat source mixes \( G_2 \):

\[
A = B_3 \times G_2, \quad \mathcal{A} = B_3 \times G_2 \neq \emptyset.
\] (7)

where the B3 scheme covered all buildings, distinguished by usage
types, and \( G_2 \) corresponded only to the buildings using electricity
for heating, distinguished by the connectivity to DH. Accordingly,

Table 1
Annual efficiency coefficients for various heating systems in the Swedish climate.

<table>
<thead>
<tr>
<th>Code</th>
<th>Type of heat source</th>
<th>Seasonal performance factor, SPF</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH</td>
<td>District heating</td>
<td>1.0</td>
<td>[6]</td>
</tr>
<tr>
<td>HO</td>
<td>Heating oil</td>
<td>0.85</td>
<td>[6]</td>
</tr>
<tr>
<td>NG</td>
<td>Natural gas</td>
<td>0.9</td>
<td>[6]</td>
</tr>
<tr>
<td>BF</td>
<td>Biofuel (firewood, pellets and other)</td>
<td>0.75</td>
<td>[6]</td>
</tr>
<tr>
<td>ER</td>
<td>Electric radiators (water, direct and air)</td>
<td>1.0</td>
<td>[6]</td>
</tr>
<tr>
<td>GSHP</td>
<td>Ground-source heat pumps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAHP</td>
<td>Exhaust air-source heat pumps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASHPW</td>
<td>Air-source heat pumps (air-water)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Grouping schemes for types of heat sources (Hn) developed to generate various
building segmentation schemes by heat source types in the electric heating case.

<table>
<thead>
<tr>
<th>H3</th>
<th>H6</th>
<th>H8</th>
<th>H13 (initial)</th>
<th>Heat source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH</td>
<td>DH</td>
<td>DH</td>
<td>DH</td>
<td>District heating</td>
</tr>
<tr>
<td>FBH</td>
<td>FBH</td>
<td>HD</td>
<td>HO</td>
<td>Heating oil</td>
</tr>
<tr>
<td>NG</td>
<td>NG</td>
<td>Natural gas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF</td>
<td>BFF</td>
<td>Biofuel (firewood)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BFH</td>
<td>FBH</td>
<td>Biofuel (pellets)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GH</td>
<td>ER</td>
<td>ER</td>
<td>ERD</td>
<td>Electric radiators (direct)</td>
</tr>
<tr>
<td>GSHP</td>
<td>GSHP</td>
<td>GSHP</td>
<td>ERA</td>
<td>Electric radiators (air)</td>
</tr>
<tr>
<td>EAHP</td>
<td>EAHP</td>
<td>EAHP</td>
<td>EAHP</td>
<td>Exhaust air-source heat pumps</td>
</tr>
<tr>
<td>ASHP</td>
<td>ASHP</td>
<td>ASHP</td>
<td>ASHP</td>
<td>Air-source heat pumps (air-air)</td>
</tr>
<tr>
<td>ASHP</td>
<td>ASHP</td>
<td>ASHP</td>
<td>ASHP</td>
<td>Air-source heat pumps (water-air)</td>
</tr>
</tbody>
</table>
all buildings not using electricity for heating ($\bar{A}$) were excluded from the study. In the case of each archetype, the information regarding EH-based heat provision was split by EH solutions of the H6 scheme. Therefore, in total there were $[B3 \times C2] = 2 \cdot 3 = 6$ archetypes, but with additional detailing for four EH solutions (ER, GSHP, EAHP, ASHP). Hereafter, this scheme is referred to as B3G2H6.

3.5. Characterisation

Each archetype subset was analysed for outliers [62], which were removed in the event of low confidence ($p < 0.01$) for the heating energy use intensity ($EH_{\text{heat}}$, kWh/m² yr). This check targeted only those records in EPC data with a high probability of error, in order to ensure that there would be no significant impact from extreme outliers. Additional outlier removal was then conducted through applying the low confidence threshold ($p < 0.01$) to energy signature parameters $\theta, c, \beta$ obtained in the modelling stage per archetype subset. This move allowed erroneous meterings (usually caused by sensor faults) to be efficiently pinpointed and the impact of the respective statistical models on the aggregated archetype models to be eliminated.

In both cases, the archetypes obtained were characterised through a deterministic approach using the data from all buildings in each archetype subset.

In Case 1 (building retrofitting), the virtual archetype building for each of the target archetypes was constructed using weight averaging of the corresponding archetype subset data by $A_{\text{temp}}$.

$$\text{feature}_{\text{archetype}} = \frac{\sum_{\text{building} \in \text{archetype}} \text{feature}_{\text{building}}}{A_{\text{temp subset}}} \quad i = 1..N \quad (8)$$

The additional assumptions for the parameters not present in EPC data and required for building energy simulations (window-to-wall ratio, U-values, standardised use) were made according to the reference data.

In Case 2 (electric heating), archetype subsets were described providing average and total values for the relevant parameters (amount of buildings, heated area $A_{\text{temp}}$, absolute and specific energy, and peak power).

Besides that, heated area $A_{\text{temp}}$ values reported in EPC data were split up proportionally to the amount of heat per source of heating $j$ for each building $i$:

$$\forall i \in \text{buildings} \quad A_{i, \text{source}_j} = \frac{Q_{\text{source}_j}}{Q_{\text{total}}} \times A_{i, \text{temp}} \quad (9)$$

so that:

$$A_{i, \text{temp}} = \sum_{\forall j} A_{i, \text{source}_j}$$

Virtual areas corresponding to the share of area heated with particular source $j$ ($A_{\text{source}_j}$) were further used to estimate heating power demand per source through the energy signature models obtained. Thus for buildings using more than one type of source for heating, the power demand estimation was based on scaling energy signature output proportionally to the share of this source in the reported total heat energy use. The peak heating power demand of each building type (scheme B3) was then estimated through applying the energy signature models (1) for the temperature of $-21^\circ C$. Finally, reference average values for each type of electric heating solution involved and average buildings in each archetype were obtained through averaging the corresponding totals by virtual heated areas and number of buildings, respectively.

3.6. Building energy simulations

Building energy simulations of virtual archetype buildings were used to estimate the effect of various energy conservation measures on the energy performance of the buildings analysed. First, a building energy simulation model was created for each building archetype in the computer simulation tool DesignBuilder. The model was set up to the characterisation obtained (Section 3.5). The simulation model obtained was then calibrated against the energy signature for the archetype building $ES_{\text{archetype}}$, that was derived from the buildings of the archetype, using (2).

Second, building retrofitting packages (one or several retrofitting measures) that are applicable for buildings of each archetype analysed were identified.

Third, the virtual archetype building models obtained were used as baseline to develop derivative models simulating the updated building energy performance due to implementation of one of the selected retrofitting packages.

Finally, the new energy signatures for each virtual archetype building in each retrofitting package were used to perform a scale-back, to update the energy signature for each building in the subset archetype, as exemplified for the balance point temperature parameter:

$$\theta_{\text{building}} \times \frac{\theta_{\text{archetype}}}{\theta_{\text{building}}} \quad \forall \text{building} \in \text{archetype}_i, i = 1..N, \quad (10)$$

where $\theta_{\text{archetype}}$ and $\theta_{\text{building}}$ are, respectively, baseline balance point temperature of the virtual archetype building and the modelled building; and $\theta_{\text{archetype}}$ and $\theta_{\text{building}}$ are, respectively, estimated balance point temperature of the virtual archetype building and the modelled building after implementation of the retrofitting package.

---

6 This temperature level is the lowest 5-day average temperature in Stockholm in 1981–2010, which by current regulations should be used to account for the peak amount of secured heat supply.
3.7. Aggregation

In Case 1 (building retrofitting), the aggregation involved calculation of specific savings per building for each retrofitting package through differentiating the specific heat demand at the baseline and after the particular retrofitting package was applied; and the total savings per package and archetype through direct summation of the absolute savings of all buildings in the archetype set.

In Case 2 (electric heating), the amount of buildings, total heated area \( A_{\text{temp}} \) and the total heat energy demand per archetype in the segmentation schemes considered (\( B3, P(H3), P(H6) \) and \( B3G2H6 \)) were calculated to make a choice about the segmentation scheme to be applied. Afterwards, electrical energy and power demand aggregations were obtained through inverse application of the SPF matrices for each type of heat source involved (Table 1). Both energy and power estimates were projected into corresponding total annual demand of electrical energy and peak power demand.

4. Results

The data-driven building archetypes approach was tested on the building stock of Stockholm for the two cases described in Section 2.2. Statistical modelling of the heating demand from the FDH data was required for the purpose of both cases, and is therefore presented separately. The further workflow was different in the two cases, but it was based on the same EPC data, comprising 30 472 energy declarations as a representation sample for the whole city building stock.

4.1. Statistical modelling of heat demand

The whole DH network of Stockholm Exergi is represented by 15 068 DH metering points. After filtering by the scope of Stockholm municipality, 12 938 metering points were left. The hourly meterings of the heat energy consumption at these metering points were used to model the heat energy demand, as described in Section 3.2. Finally, after performing the linkage of EPC and FDH data and dropping the outliers, 6732 metering points were left, and were used for further analysis in both cases. The accuracy of the energy signature models developed was assessed using the quality indices depicted in Fig. 4.

As can be seen from Fig. 4, the majority of the energy signature models obtained had an acceptable magnitude of error, e.g. under ASHRAE Guideline 14 [63] MASE, NRMSE and \( R^2 \) should meet the criteria \( \pm 10\% <30\% >0.75 \), respectively. Hence, it was reasonable to use these models to calibrate and estimate the specific heat demand of archetype buildings constructed in both cases.\(^7\)

4.2. Case 1. Building retrofitting

Three building archetypes were developed for the building retrofitting case (Table 4) as described in Section 3.4.\(^8\)

Archetype \( A_1 \) represents multi-residential buildings constructed in 1946–1975, corresponding to the two consecutive periods of developing affordable housing in Sweden: Folkhemsbygget (1946–1960) and Miljonprogrammet (1961–1975). This is therefore the most widespread type of building not only in Stockholm, but generally in Sweden. Archetype \( A_2 \) represents all office buildings and was chosen because it is another widespread type of building, having a noticeably different type of activities compared with \( A_1 \) buildings. Archetype \( A_3 \) represents recently (1996–…) built multi-residential buildings, and was chosen as it reflects the state-of-the-art of Swedish building construction. New buildings are often used to benchmark the energy performance of older buildings. At the same time, the potential for energy saving in new buildings can indicate the efficiency of the current building code and further need for improvement. Finally, all three archetypes were limited to buildings using DH as their main source of heat, to use the energy signatures obtained in Section 4.1. The statistical distribution of the features further used for construction of the virtual archetype buildings is depicted in Fig. 5. A detailed summary is provided in Appendix A (Table).

Virtual archetype buildings were created in DesignBuilder from the reference data and the aggregated features of the buildings characterised from each archetype subset according to (2).

After the baseline simulation models were set up, they were calibrated using the aggregated energy signature of each archetype (\( \mathbf{t}_0 \) in Table 5) obtained through averaging the energy signature model parameters (2). The calibrated dynamic simulation models

\(^7\) Tukey boxplots are heavily used in this paper to communicate distributions. Boxes represent 25th percentile, median estimate and 75th percentile. Whiskers represent minimums and maximums for 1.5 IQR (inter-quartile range). For more details, see e.g. Ref. [74].

\(^8\) Hereafter, metering points are limited to those used for calibration purposes, with a verified linkage with the buildings analyzed and providing heat to buildings of only one archetype.
of the virtual archetype buildings were then updated to model different retrofitting packages selected for analysis for each archetype (Table 6).

The updated energy signatures of the virtual archetype buildings (Table 5) were used to project the modified energy performance of each building in the different archetype subsets for each retrofitting package analysed, applying the scaling technique (4).

Fig. 6 depicts the estimates of energy savings obtained for various building energy retrofits, aggregated per archetype subset and retrofitting package applied.

As can be observed from Fig. 6a, the specific savings for a particular retrofitting package varied depending on the archetype applied (heat recovery ventilation in A1 and A3). In the case of combined packages (r101-r103), the total saving effect was sometimes noticeably less than that of the same measures individually, which confirmed previous findings [64]. The total absolute savings for buildings in three archetypes (Fig. 6b) were estimated at 738 GWh/year.

Table 5
Energy signatures of virtual archetype buildings corresponding to calibrated baseline (r0) or application of selected retrofitting packages.

<table>
<thead>
<tr>
<th>Archetype</th>
<th>Retrofitting package</th>
<th>$\bar{t}$, balance temperature, °C</th>
<th>$c$, base load, W/m$^2$</th>
<th>$\beta$, performance coefficient, W/m$^2$.K</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>$r_0$, baseline</td>
<td>15.32</td>
<td>4.45</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td>$r_1$, HRV</td>
<td>15.22</td>
<td>4.41</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>$r_2$, EEW</td>
<td>15.25</td>
<td>4.45</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>$r_{101}$, HRV + EEW</td>
<td>15.39</td>
<td>4.45</td>
<td>1.10</td>
</tr>
<tr>
<td>A2</td>
<td>$r_0$, baseline</td>
<td>13.29</td>
<td>1.29</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>$r_2$, EEW</td>
<td>11.47</td>
<td>1.07</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>$r_3$, IT21</td>
<td>13.22</td>
<td>1.41</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>$r_4$, VFR</td>
<td>13.09</td>
<td>1.35</td>
<td>0.79</td>
</tr>
<tr>
<td>A3</td>
<td>$r_0$, baseline</td>
<td>13.63</td>
<td>3.10</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>$r_1$, HRV</td>
<td>12.13</td>
<td>3.12</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>$r_2$, EEW</td>
<td>14.13</td>
<td>3.11</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>$r_3$, IT21</td>
<td>11.86</td>
<td>3.12</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>$r_{103}$, IT21 + EEW + HRV</td>
<td>10.63</td>
<td>3.06</td>
<td>0.57</td>
</tr>
</tbody>
</table>
The segmentation was conducted in two stages. First, the building stock was split into three non-overlapping groups of buildings (hereafter B3): S = single family dwellings, M = multi-residential buildings, and O = other buildings (offices, schools, hospitals, etc.), as illustrated in Table 3. This stage of segmentation was performed using the tax agency (Skatteverket) building categories (see Table A2 in Appendix A for full details). This segmentation scheme was chosen as it is well-established and used by many researchers (e.g. Ref. [65] and practitioners (e.g. Ref. [66]) and as it provided direct consistency with other energy-related studies performed by the City of Stockholm.

Second, the building stock was further split by the heat source types. The initial EPC data contained information about the heat sources distinguished into 13 types, and therefore several grouping schemes were applied in the further analysis depending on the question addressed (Table 2). The H13 scheme was simplified to a H8 scheme through combining all electric radiators, air-source heat pumps and biofuel-based heaters into three overarching groups. The H3 scheme was then set up as the most general, providing a split of the heat sources into three main types (DH = district heating, FBH = fossil- and biofuel-based heating, and EH = electric heating), which provided a general overview of the building stock by total amount of heat and the $\mathcal{P}(H3)$ mixes of used heat sources (Fig. 7a). However, further application of the $\mathcal{P}(H8)$ scheme revealed an overwhelmingly large variety ($n = 77$) of mixes of heat sources. Therefore, to reduce this complexity, the H6 scheme was constructed and further applied to analyse the most typical mixes of heat sources (Fig. 7b). It provided a reasonable compromise for analysis of mixes through keeping more details for the EH heating solutions that were the focus of this study, while all other types of heat sources were simplified to DH and fuel-based burners (FBH).

The total number of archetypes that would derive from using $\mathcal{P}(H3)$ or $\mathcal{P}(H6)$ in the whole segmentation would be too large; in the case of $\mathcal{P}(H3)$ the upper limit on the number of resulting archetypes was $\mid B3 \times \mathcal{P}(H3) \mid = 3 \cdot (2^3 - 1) = 3.7 = 21$ archetypes, and for the case of the $\mathcal{P}(H6)$ it was accordingly $\mid B3 \times \mathcal{P}(H6) \mid = 3 \cdot (2^6 - 1) = 3.63 = 189$ archetypes. While such microdivisioning could still be relevant to shape policies and incentives targeting very particular narrow segments, that was clearly not practical from the perspective of this study.

As can be seen from the structure of the heating by heat sources (Fig. 7), DH plays a dominant role in provision of heating in Stockholm, being associated with ~85% of the total city heat demand. Even though electric heating represents only 10.5% of the heat demand, city-wise this is still a considerable amount of energy (721 GWhheat or 383 GWh Electro), which is currently the subject for further growth due to further efficiency improvement of heat pumps and the relatively low price of electricity. The graphic illustration of the mixes also highlighted the marked difference in the role of EH-based heat sources for sole EH usage (one or several EH-based solutions) or mixed EH usage, when EH solutions are used in combination with DH.

Therefore, two main target groups, G1 (“EH mainly”) and G2 (“EH partly”), were identified from the space of mixes of types of heating sources $\mathcal{P}(H3)$. The whole segmentation scheme was then constructed as a Cartesian product of three building types (Table 3) and these two groups as described in Section 3.4, resulting in the scheme B3G2H6. The counts of buildings per archetype/EH solution are shown in Table A3 in Appendix A.

### 4.3. Case 2. Electric heating

The segmentation was conducted in two stages.

First, the building stock was split into three non-overlapping groups of buildings (hereafter B3): S = single family dwellings, M = multi-residential buildings, and O = other buildings (offices, schools, hospitals, etc.), as illustrated in Table A2 in Appendix A for full details. This segmentation scheme was chosen as it is well-established and used by many researchers (e.g. Ref. [65] and practitioners (e.g. Ref. [66]) and as it provided direct consistency with other energy-related studies performed by the City of Stockholm.

Second, the building stock was further split by the heat source types. The initial EPC data contained information about the heat sources distinguished into 13 types, and therefore several grouping schemes were applied in the further analysis depending on the question addressed (Table 2). The H13 scheme was simplified to a H8 scheme through combining all electric radiators, air-source heat pumps and biofuel-based heaters into three overarching groups. The H3 scheme was then set up as the most general, providing a split of the heat sources into three main types (DH = district heating, FBH = fossil- and biofuel-based heating, and EH = electric heating), which provided a general overview of the building stock by total amount of heat and the $\mathcal{P}(H3)$ mixes of used heat sources (Fig. 7a). However, further application of the $\mathcal{P}(H8)$ scheme revealed an overwhelmingly large variety ($n = 77$) of mixes of heat sources. Therefore, to reduce this complexity, the H6 scheme was constructed and further applied to analyse the most typical mixes of heat sources (Fig. 7b). It provided a reasonable compromise for analysis of mixes through keeping more details for the EH heating solutions that were the focus of this study, while all other types of heat sources were simplified to DH and fuel-based burners (FBH).

The total number of archetypes that would derive from using $\mathcal{P}(H3)$ or $\mathcal{P}(H6)$ in the whole segmentation would be too large; in the case of $\mathcal{P}(H3)$ the upper limit on the number of resulting archetypes was $\mid B3 \times \mathcal{P}(H3) \mid = 3 \cdot (2^3 - 1) = 3.7 = 21$ archetypes, and for the case of the $\mathcal{P}(H6)$ it was accordingly $\mid B3 \times \mathcal{P}(H6) \mid = 3 \cdot (2^6 - 1) = 3.63 = 189$ archetypes. While such microdivisioning could still be relevant to shape policies and incentives targeting very particular narrow segments, that was clearly not practical from the perspective of this study.

As can be seen from the structure of the heating by heat sources (Fig. 7), DH plays a dominant role in provision of heating in Stockholm, being associated with ~85% of the total city heat demand. Even though electric heating represents only 10.5% of the heat demand, city-wise this is still a considerable amount of energy (721 GWhheat or 383 GWh Electro), which is currently the subject for further growth due to further efficiency improvement of heat pumps and the relatively low price of electricity. The graphic illustration of the mixes also highlighted the marked difference in the role of EH-based heat sources for sole EH usage (one or several EH-based solutions) or mixed EH usage, when EH solutions are used in combination with DH.

Therefore, two main target groups, G1 (“EH mainly”) and G2 (“EH partly”), were identified from the space of mixes of types of heating sources $\mathcal{P}(H3)$. The whole segmentation scheme was then constructed as a Cartesian product of three building types (Table 3) and these two groups as described in Section 3.4, resulting in the scheme B3G2H6. The counts of buildings per archetype/EH solution are shown in Table A3 in Appendix A.

---

**Table 6**

<table>
<thead>
<tr>
<th>Retrofitting package</th>
<th>Target archetypes</th>
<th>Involved measures</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>f0 (baseline)</td>
<td>A1, A2, A3</td>
<td>–</td>
<td>Calibrated baseline</td>
</tr>
<tr>
<td>f1</td>
<td>A1, A3</td>
<td>HRV</td>
<td>Heat recovery ventilation (FX + HP)</td>
</tr>
<tr>
<td>f2</td>
<td>A1, A2, A3</td>
<td>EEW</td>
<td>Energy efficient windows (U = 0.78 W/m2K)</td>
</tr>
<tr>
<td>f3</td>
<td>A2, A3</td>
<td>IT21</td>
<td>Adapt indoor temp to 21 °C</td>
</tr>
<tr>
<td>f4</td>
<td>A2</td>
<td>VFR</td>
<td>Adapt ventilation flow rate to min (0.35 l/s/m2)</td>
</tr>
<tr>
<td>f101</td>
<td>A1</td>
<td>HRV + EEW</td>
<td>Heat recovery ventilation + energy efficient windows</td>
</tr>
<tr>
<td>f102</td>
<td>A2</td>
<td>IT21 + EEW + VFR</td>
<td>Adapt indoor temperature + energy efficient windows + adapt ventilation flow rate to min</td>
</tr>
<tr>
<td>f103</td>
<td>A3</td>
<td>IT21 + EEW + HRV</td>
<td>Adapt indoor temperature + energy efficient windows + heat recovery ventilation</td>
</tr>
</tbody>
</table>

---

**Fig. 6.** Specific (a) and absolute (b) heat savings of the different retrofitting packages per selected archetype.
To better understand the character of the load covered by separate types of EH heat sources, shares of particular solutions \(\frac{Q_{\text{source}}}{Q_{\text{total}}}\) were calculated for each building per archetype (Fig. 8). This graph proved to be particularly useful in representing the technological differences between the solutions. The width of the interquartile range (IQR) can be treated as a characteristic for the diversity of the installed systems, ranging from rather narrow ranges of 5–20% width (e.g. GSHP in G_{I} or ER in G_{II}) to basically universal ranges of more than 60% for such general sets of heat sources as FBH or EH. Comparing the distributions of shares group-wise further demonstrated the significant difference between the two target groups, e.g. there was a clear distinction for the modes using GSHP depending on the availability of DH (G_{I} vs. G_{II}). Furthermore, this allowed the systems used to be classified by the type of load addressed, e.g. GSHP in G_{I} corresponds to the base load, ASHP in both G_{I} and G_{II} can be attributed to intermediate load, and ER in G_{II} corresponds to the peak load. While these observations are not of significant novelty as information about these general types of heat sources, they can clearly be useful for justification of modelling assumptions in energy studies on a large scale (e.g. district or urban).

Following the B3G2H6 scheme, the total heat energy demand per archetype was calculated with further disaggregation by types of EH solutions (Table 7).

Applying the respective SPF coefficients (Table 1) to the annual heat values (Table 7) backwards, estimates for the annual amounts of electrical energy consumed for heating purposes were obtained (Table 8).

As follows from Table 8, the electricity demand for the building stock is 81% for G_{I} and 19% for G_{II}, where large buildings (multi-residential & other) constitute the majority in the group of ‘potential switchers’ (G_{II}, EH partly). It can be also seen that electric radiators (ER), which usually require relatively low capital investments, are responsible for more than half of the total energy demand.
To obtain the estimates for peak power demand, the energy signature models of buildings using DH only ($DH_\text{(EH_FBH)}$) were used (Section 4.1). They were subset and aggregated per building type in B3 (S = single-family, M = multi-residential, O = other) following (2). The parameters for resulting energy signatures per building type are given in Table 9. Virtual areas corresponding to the share of area heated with a particular source were then used to estimate the heat power demand per heating solution in each building according to (3). Heating solutions are structured by the scheme H6 (DH – district heating, FBH – fossil- and biofuel-based, EH – electric, split into ER – electric radiators, GSPH – ground-source heat pumps, EAHP – exhaust air-source heat pumps, and ASHP – air-source heat pumps).

Finally, the counts obtained for required electricity demand and the estimates for peak electric power required for electric heating were used to provide average peak electric power per building (Table 11), average peak electric power per area (Table 12) and average annual specific energy demand (Table 13).

The analysis allowed the lower margin for peak electric power demand for electric heating to be estimated as 147 MW. Of this, 119 MW corresponded to $G_1$ (EH mainly) and 35 MW to $G_2$ (EH partly), i.e., the group of buildings connected to the DH network, as confirmed by non-zero consumption of heat attributed to DH. While this power is clearly not enough to compensate for the expected deficit for power capacity (Fig. 1), it has potential to provide additional flexibility to the urban energy system of Stockholm.

### 5. Discussion

The exponential growth in the availability and quality of building energy data, along with the continuous development of computational infrastructures and analytical methods, sets a welcoming environment for further expansion of UBEM-based tailor-made solutions to address various urban energy challenges, utilising as much information as possible for each problem context. This drives further improvements in the quality of the models developed, and, at the same time, decreases tolerance to simplifications implied by the universal archetyping schemes. The results obtained in this study suggest that any urban building energy modelling exercise should recognise the need for alignment of the chosen archetyping scheme with the problem to be solved, data provided, methods used and project resources available.

Until recently, archetypes were mainly developed as universal reference input data for individual building energy models (BEMs) until recently, archetypes were mainly developed as universal reference input data for individual building energy models (BEMs) until recently, archetypes were mainly developed as universal reference input data for individual building energy models (BEMs) until recently, archetypes were mainly developed as universal reference input data for individual building energy models (BEMs) until recently, archetypes were mainly developed as universal reference input data for individual building energy models (BEMs).
and the earliest UBEMs [26,67]. However, the topicality of such building typologies is decreasing with the growing availability of data and the recent developments in UBEM [79]. As more rich datasets emerge, the issue of similarity from a building energy

### Table 10

Peak electrical power demand associated with electric heat (EH) in buildings segmented by the archetyping scheme B3G2H6, MWel.

<table>
<thead>
<tr>
<th>BJ</th>
<th>EH</th>
<th>GI – EH:DH (EH mainly)</th>
<th>GII – EH:DH (EH partly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>48.8</td>
<td>48.5</td>
<td>30.8</td>
</tr>
<tr>
<td>M</td>
<td>51.8</td>
<td>30.9</td>
<td>17.1</td>
</tr>
<tr>
<td>O</td>
<td>46.3</td>
<td>32.4</td>
<td>25.4</td>
</tr>
</tbody>
</table>

### Table 11

Average peak power per building in each segment of the archetyping scheme B3G2H6, kWel/building.

<table>
<thead>
<tr>
<th>BJ</th>
<th>EH</th>
<th>GI – EH:DH (EH mainly)</th>
<th>GII – EH:DH (EH partly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>9.1</td>
<td>4.7</td>
<td>4.3</td>
</tr>
<tr>
<td>M</td>
<td>51.4</td>
<td>27.0</td>
<td>34.8</td>
</tr>
<tr>
<td>O</td>
<td>71.4</td>
<td>33.6</td>
<td>31.8</td>
</tr>
<tr>
<td>AVG</td>
<td>21.2</td>
<td>9.0</td>
<td>8.7</td>
</tr>
</tbody>
</table>

### Table 12

Average peak power in buildings of the archetyping scheme B3G2H6, W/m².

<table>
<thead>
<tr>
<th>BJ</th>
<th>EH</th>
<th>GI – EH:DH (EH mainly)</th>
<th>GII – EH:DH (EH partly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>27.5</td>
<td>27.5</td>
<td>47.9</td>
</tr>
<tr>
<td>M</td>
<td>26.8</td>
<td>27.3</td>
<td>50.5</td>
</tr>
<tr>
<td>O</td>
<td>28.6</td>
<td>32.0</td>
<td>45.4</td>
</tr>
<tr>
<td>AVG</td>
<td>27.6</td>
<td>28.6</td>
<td>47.6</td>
</tr>
</tbody>
</table>

### Table 13

Average electrical energy use intensity for the electric heating in each segment of the archetyping scheme B3G2H6, kWhel/m².yr

<table>
<thead>
<tr>
<th>BJ</th>
<th>EH</th>
<th>GI – EH:DH (EH mainly)</th>
<th>GII – EH:DH (EH partly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>85.0</td>
<td>84.9</td>
<td>123.9</td>
</tr>
<tr>
<td>M</td>
<td>64.0</td>
<td>72.9</td>
<td>112.1</td>
</tr>
<tr>
<td>O</td>
<td>67.2</td>
<td>78.3</td>
<td>106.1</td>
</tr>
<tr>
<td>AVG</td>
<td>71.9</td>
<td>79.7</td>
<td>114.8</td>
</tr>
</tbody>
</table>

---

Fig. 9. Structure of the peak electrical power demand (corresponding to the electric heating at −21°C) according to the archetyping scheme B3G2H6. More saturated segments above correspond to GI (EH mainly), less saturated segments below to GII (EH partly).
performance perspective is gaining more importance, as it allows data-driven estimation of building energy performance to be performed for different building stocks analysed using individual building-centric clusters of similar buildings [24, 25]. This would eliminate the subjectivity of modellers and increase the scalability and transferability of the results obtained. Furthermore, on-demand segmentation would provide the possibility to focus on those aspects of building energy performance that are relevant for a particular problem, improving the efficiency of the archetypes derived from the perspective of complexity and interpretability of modelling outcomes.

The developments in this study allowed the following criteria for building archetypes to be formulated:

- **Quality** — amount of error induced by the archetype-caused simplification.
- **Computational complexity** — structural and computing complexity associated with the archetypes developed.
- **Intensity** — redundancy for the archetyping scheme given the requirements on the levels of quality and/or detail.
- **Expendability** — flexibility in regulating the level of detail for the archetyping scheme.
- **Data demand** — amount and type of input data required.
- **Data robustness** — sensitivity to data anomalies.
- **Interpretability** — understanding for energy performance of the building stock.
- **Representativeness** — representation for all aspects of the problem addressed.
- **Transferability** — possibility of applying the archetypes developed in other cases.
- **Completeness** — coverage of the building stock.
- **Customisability** — possibility of user extension of archetypes depending on the problem addressed.

Many of the above-mentioned criteria were relevant for both cases in this study. In particular, in terms of data demand, spatial or building geometry data were not needed for input in both cases. It was more important to monitor the quality of archetyping in Case 1 (building retrofitting), while this was not recognised as a problem for Case 2 (electric heating), which had a limited amount of uncertainty added to data by the archetyping process. With regard to expandability, flexibility in regulating the level of detail for the archetyping scheme was not possible for archetypes in Case 1, while Case 2 had this feature by design. Conversely, transferability was not feasible for the results of archetyping in Case 2, while in Case 1 the results of archetyping could definitely be transferred to other Swedish cities. More importantly, the archetyping approaches used in both cases are transferrable to any other Swedish city.

In both cases, the chosen segmentation (categories and their granularity) was the result of a compromise between the level of detail and generalisation considered appropriate for the problem analysis, as demonstrated in previous studies [12, 22].

In Case 1 (building retrofitting), the archetyping aimed to obtain archetype subsets that would be not only sufficiently large, but also sufficiently homogeneous to ensure feasibility of the proposed retrofitting measures and to allow for justification of assumptions about typical activities and related internal gains from the reference data. Furthermore, the set of buildings analysed was limited only to those connected to the district heating network, as this allowed building energy performance of all buildings analysed to be statistically modelled using the measured data on actual heat energy use.

In Case 2 (electric heating), the segmentation aimed to cover the whole city building stock and was more general, as further splits could be performed based on the source of heat. The particular segmentation schemes were used according to the needs of Stockholm City energy planning unit and were largely required to preserve compliance with previous and adjacent studies. At the same time, the approach presented here allows full flexibility in setting up any other segmentation scheme that can be expressed through conditioning of features present in energy declarations, allowing instant update of all results obtained.

Limitations of this study concern the uncertainty embedded in the input data, statistical models and approaches used for characterisation and peak power estimation. The EPC data used as core input data for both cases have proven to be an important source of information about building stock, especially in Sweden. Though the data contain some errors [46, 68], to our knowledge they are the best available data source on building energy on the urban scale in Sweden. However, there are two peculiarities of this data source that affected the results in Case 2. First, the national regulations for buildings to have an energy audit are not as demanding in the case of single-family houses, which resulted in significant under-representation of this segment in the energy and power calculations performed. According to the national statistics agency, Stockholm has 42,867 single-family buildings [69], but only 13,123 were represented in the input data. Second, the energy audits, which are the only input for the EPC register, are required to be conducted once a decade, and therefore a noticeable temporal lag is present in the results for Case 2. For example, according to EPC data, a recently built environmental district Hammarby Sjöstad has 59 housing associations with only 1 GSHP, 2 EAHPs and 0 ASHPs installed, while in reality there are confirmed installations for 7 GSHPs, 19 EAHPs and 16 ASHPs [70]. This is in line with the general trend of increased penetration of heat pump technology [2].

Regarding the statistical models, it should be noted that, even though energy signature models embed end uses and thermal mass of buildings [4], they do not allow these effects, which obviously are different per building analysed, to be isolated and estimated. Besides, some particular deviations related to peak frosts and base summer loads could be observed for individual buildings, which corresponds to previous observations [12].

As for characterisation, even though a deterministic approach is known to be less precise than a probabilistic approach [19], it was assumed to be sufficient due to: a) further calibration and scaling of models with energy signatures based in Case 1; and b) purely exploratory and compression function of archetypes in Case 2, which clearly could be characterised probabilistically if further required as a model input.

The results obtained in Case 1 (building retrofitting) do not account for the rebound effect and energy performance gap that has been observed after introducing retrofitting packages [31]. To obtain a data-driven estimate of this phenomenon, an additional validation study based on measured data for buildings undergoing retrofitting projects could be performed [71].

The estimation of peak power in Case 2 (electric heating) was based on the assumption that buildings using EH have on average the same heat load profiles as those connected to DH, which allowed the statistical models of heat load profiles to be applied for different building usage types (scheme B3: S = single-family, M = multi-residential, O = other) derived from FDH data. To avoid such uncertainty, the verification study could be performed using the high-resolution metering data of building electricity usage structured by types of heating solutions. However, neither applied nor proposed approaches would account for application of heating appliances embedded in regular household loads.

It is worth noting that the cooling demand was beyond the scope of the study. Stockholm has a developed district cooling
system [65] and the cooling demand can be modelled with energy signatures fitted by actual energy use in the same way it was done for heat. However, the possible double use of heat pumps and the low granularity of the information about cooling sources in EPC data made it scarcely possible to estimate the cooling demand covered by electricity.

As can be observed throughout this paper, mean values and sample distributions remain the main form of aggregating building stock information for the purpose of simplifying the models. However, median and IQR should also be considered for representation purposes in the case of skewed distributions, as they are less sensitive to outliers. Therefore Tukey boxplots were used in a number of diagrams in this paper.

The spatial perspective was beyond the scope of this paper. However, it is another dimension of analysis that is relevant for energy planners, especially for Case 2. To explore the potential synergies between the local power distribution grid and DH pipe network, their topologies would need to be linked. This would allow existing surpluses and deficits of capacity with the EH-related loads identified in this study to be mapped and an action plan for local stakeholders to be developed.

Sweden already has precedent, as various policies have been introduced to stimulate conversion to energy sources for space heating other than electricity [72]. This study did not seek to identify any solution for the power capacity problem. Instead, it aimed to highlight the additional potential for flexibility embedded in electric heating that could be utilised through more intelligent orchestration of energy in cities involving technologies and policies developed jointly by government, grid operators, energy utilities and cities.

6. Conclusions

The study demonstrated how using rich datasets allows different building archetypes to be developed, depending on the urban energy challenges addressed. Two cases were analysed for Stockholm based on energy performance certificates (EPC) and measured data. In the case of building energy retrofitting, three building archetypes (multi-residential buildings 1946–1975, offices and multi-residential buildings 1996–2011) were developed, which allowed energy savings for seven retrofitting packages to be estimated. In the case of electric heating (EH), six archetypes were developed as a combination of three usage types (single-family houses, multi-residential buildings and other) and two groups of buildings using EH by heat sources (using DH and not using DH), with details for four main types of EH solutions (electric radiators, ground-, exhaust air- and air-source heat pumps). The identified lower bound for peak electric power for EH (147 MW) could be addressed to improve power capacity of the local electric grid.

The noticeable improvement in UBEM workflows demonstrated in this paper derives from the use of the EPC data linked with measured data, accompanied by data quality control. It allowed statistical models to be fitted prior to the segmentation, independent of the choice of archetype segmentation scheme, and two different urban energy challenges to be explored using the same input data, as demonstrated with the two cases analysed. The growing availability of data greatly decreases the amount of assumptions required about building parameters in analysis or simulations, which can now rely largely on actual information about individual buildings. However, this imposes new challenges related to data quality and consistency, along with more intelligent automatic segmentations that are required to provide computational scalability of constructed UBEMs.

The cases for building retrofitting and electric heating analysed here exemplify how the emergence of linked building energy data gradually changes the paradigm of UBEM itself, driving further progression from stand-alone energy datasets with single-logic models to urban data lakes fuelling on-request multiple-purpose data intelligence services.

Acknowledgements

This research was supported by the E2B2 research programme run by IQ Samhällsbyggnad and funded by the Swedish Energy Agency (project No. 40846–1). The study was performed through collaboration between the authors, the City of Stockholm, Hammarby Sjöstad Organisation of Housing Associations (Sjöstadsföreningen), Lassila & Tikanoja, the Association of Property Owners (Fastighetsägarna) and AB Stockholm Exergi.

We express gratitude to Boverket (The Board of Housing, Building and Planning) for providing extracts from EPC database “Gripen”. We also thank Mary McAfee for professional language check.

Appendix A

This section provides additional details for:

- Virtual archetype buildings constructed for the building retrofitting case (Table A1)
- Aggregation of the tax building categories into three building types (scheme B3) used in the electric heating case (Table A2)
- Number of buildings per building archetype in the B3G2 segmentation scheme in the electric heating case (Table A3)
- Distribution of virtual areas for EH-based solutions in the B3G2 segmentation scheme in the electric heating case (Table A4)

Table A1

<table>
<thead>
<tr>
<th>Archetype/Feature</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>IQR</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUI, kWh/m².yr</td>
<td>170.4</td>
<td>167</td>
<td>32.7</td>
<td>38</td>
<td>41</td>
<td>480</td>
</tr>
<tr>
<td>Heated area Atemp,m²</td>
<td>2816.2</td>
<td>2162</td>
<td>2572.8</td>
<td>2035</td>
<td>63</td>
<td>38230</td>
</tr>
<tr>
<td>Year of construction</td>
<td>1957.1</td>
<td>1955</td>
<td>8.2</td>
<td>15</td>
<td>1946</td>
<td>1975</td>
</tr>
<tr>
<td>Number of apartments</td>
<td>32.2</td>
<td>24</td>
<td>30.8</td>
<td>24</td>
<td>0</td>
<td>534</td>
</tr>
<tr>
<td>Number of floors</td>
<td>4.4</td>
<td>3</td>
<td>2.5</td>
<td>2</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Number of basement floors</td>
<td>1.0</td>
<td>1</td>
<td>1.4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Number of stairwells</td>
<td>3.1</td>
<td>3</td>
<td>2.7</td>
<td>3</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>A₀, Offices (N = 1028)</td>
<td>161.1</td>
<td>154.5</td>
<td>54.3</td>
<td>65</td>
<td>46</td>
<td>476</td>
</tr>
<tr>
<td>EUI, kWh/m².yr</td>
<td>9532.8</td>
<td>5046</td>
<td>9223</td>
<td>55</td>
<td>98</td>
<td>785</td>
</tr>
<tr>
<td>Year of construction</td>
<td>1936.5</td>
<td>1957</td>
<td>66.0</td>
<td>55.25</td>
<td>1600</td>
<td>2011</td>
</tr>
<tr>
<td>Number of apartments</td>
<td>3.1</td>
<td>0</td>
<td>12.6</td>
<td>0</td>
<td>0</td>
<td>175</td>
</tr>
</tbody>
</table>

(C) O. Pasichnyi et al. / Energy 181 (2019) 360–377
### Table A2
Stockholm building stock segmented by types of usage (B3) through tax agency (Skatteverket) building categories [73]:1961) in the electric heating case

<table>
<thead>
<tr>
<th>B3 code</th>
<th>Tax category (in Swedish)</th>
<th>Amount of buildings, N</th>
<th>Heated area $A_{\text{temp}}$, m$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S (single-family)</td>
<td>210 - Småhusenhet, tomtark till helårsbostad</td>
<td>4</td>
<td>1322</td>
</tr>
<tr>
<td></td>
<td>220 - Småhusenhet, helårsbostad för 1–2 familjer</td>
<td>12869</td>
<td>2181 241</td>
</tr>
<tr>
<td></td>
<td>221 - Småhusenhet, fritidsbostad för 1–2 familjer</td>
<td>15</td>
<td>2814</td>
</tr>
<tr>
<td></td>
<td>222 - Småhusenhet, flera småhus, sammanaget bostäder för mer än 2 familjer (helårs- eller fritidsbostad)</td>
<td>209</td>
<td>55 745</td>
</tr>
<tr>
<td></td>
<td>223 - Småhusenhet med lokaler</td>
<td>13</td>
<td>3659</td>
</tr>
<tr>
<td>M (multi-residential)</td>
<td>310 - Hyreshusenhet, tomtmark</td>
<td>13</td>
<td>45 954</td>
</tr>
<tr>
<td></td>
<td>320 - Hyreshusenhet, huvudsaklig bostäder</td>
<td>11362</td>
<td>27 877 299</td>
</tr>
<tr>
<td></td>
<td>321 - Hyreshusenhet, kiosk</td>
<td>2</td>
<td>2885</td>
</tr>
<tr>
<td></td>
<td>322 - Hyreshusenhet, parkeringshus/garage</td>
<td>11</td>
<td>7767</td>
</tr>
<tr>
<td></td>
<td>323 - Hyreshusenhet, hotell eller restaurangbyggnad</td>
<td>76</td>
<td>577 320</td>
</tr>
<tr>
<td></td>
<td>324 - Hyreshusenhet, kiosk</td>
<td>2</td>
<td>2885</td>
</tr>
<tr>
<td>O (other)</td>
<td>325 - Hyreshusenhet, parkeringshus/garage</td>
<td>11</td>
<td>7767</td>
</tr>
<tr>
<td></td>
<td>326 - Hyreshusenhet, kontor eller dylikt beläget inom industrimark</td>
<td>51</td>
<td>419 325</td>
</tr>
<tr>
<td></td>
<td>326 - Hyreshusenhet, kontor eller dylikt beläget inom industrimark</td>
<td>13</td>
<td>45 954</td>
</tr>
<tr>
<td></td>
<td>380 - Hyreshusenhet</td>
<td>1</td>
<td>1756</td>
</tr>
<tr>
<td></td>
<td>381 - Hyreshusenhet, skattefri enligt kap 4 § FTL</td>
<td>3</td>
<td>6352</td>
</tr>
<tr>
<td></td>
<td>400 - Ej fastställd typ av industrienhed</td>
<td>169</td>
<td>1 044 449</td>
</tr>
<tr>
<td></td>
<td>420 - Industrienhet, industrihotell</td>
<td>2</td>
<td>26 317</td>
</tr>
<tr>
<td></td>
<td>421 - Industrienhet, kemisk industri</td>
<td>1</td>
<td>22 346</td>
</tr>
<tr>
<td></td>
<td>423 - Industrienhet, metall- och maskinindustri</td>
<td>2</td>
<td>4163</td>
</tr>
<tr>
<td></td>
<td>426 - Industrienhet, annan tillverkningsindustri</td>
<td>8</td>
<td>63 126</td>
</tr>
<tr>
<td></td>
<td>431 - Industrienhet, reparationsverkstad</td>
<td>3</td>
<td>2359</td>
</tr>
<tr>
<td></td>
<td>432 - Industrienhet, lager</td>
<td>6</td>
<td>28 811</td>
</tr>
<tr>
<td></td>
<td>433 - Industrienhet, annan ovrig byggnad</td>
<td>1</td>
<td>4356</td>
</tr>
<tr>
<td></td>
<td>498 - Industrienhet, gata- eller parkmark</td>
<td>2</td>
<td>792</td>
</tr>
<tr>
<td></td>
<td>520 - Agarlägenshedenhet, bostad</td>
<td>3</td>
<td>2174</td>
</tr>
<tr>
<td></td>
<td>823 - Specialenhed, vårdbyggnad</td>
<td>529</td>
<td>1 260 181</td>
</tr>
<tr>
<td></td>
<td>824 - Specialenhed, bad-, sport- eller idrottsanläggning</td>
<td>63</td>
<td>309 918</td>
</tr>
<tr>
<td></td>
<td>825 - Specialenhed, skolbyggnad</td>
<td>575</td>
<td>1 888 395</td>
</tr>
<tr>
<td></td>
<td>826 - Specialenhed, kulturbyggnad</td>
<td>116</td>
<td>544 665</td>
</tr>
<tr>
<td></td>
<td>828 - Specialenhed, allmän byggnad</td>
<td>167</td>
<td>1 083 437</td>
</tr>
<tr>
<td></td>
<td>829 - Specialenhed, kommunikationsbyggnad</td>
<td>57</td>
<td>549 446</td>
</tr>
<tr>
<td></td>
<td>901 - Okand taxeringsenhed</td>
<td>1</td>
<td>595</td>
</tr>
<tr>
<td></td>
<td>999 - Okand taxeringsenhed med värde &lt;1000 kr</td>
<td>89</td>
<td>366 318</td>
</tr>
</tbody>
</table>

### Table A3
Number of buildings per segment of the B3G2 archetyping scheme in the electric heating case

<table>
<thead>
<tr>
<th>B3 code</th>
<th>TOTAL</th>
<th>$G_2 - \text{EH} \cdot \text{DH}$ (EH mainly)</th>
<th>$G_2 - \text{EH} \cdot \text{DH}$ (EH partly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>10 695</td>
<td>10 344</td>
<td>7 165</td>
</tr>
<tr>
<td>M</td>
<td>1870</td>
<td>1142</td>
<td>493</td>
</tr>
<tr>
<td>O</td>
<td>1297</td>
<td>965</td>
<td>799</td>
</tr>
<tr>
<td>TOTAL</td>
<td>13 862</td>
<td>12 451</td>
<td>8 457</td>
</tr>
</tbody>
</table>


G2: $G_2 = \text{EH}$ mainly, $G_2 = \text{EH}$ partly.

H6: ER = electric radiators, GSHP = ground-source heat pumps, EAHP = exhaust air-source heat pumps, ASHP = air-source heat pumps.
Table A4
Summary of ‘virtual areas’ (areas corresponding to the share of area heated per particular source) for EH-based solutions in the B3G2 segmentation scheme, $A_{\text{temp}}-m^2$.

<table>
<thead>
<tr>
<th>B3</th>
<th>Total</th>
<th>Sum</th>
<th>ER</th>
<th>GCHP</th>
<th>EAH</th>
<th>ASHP</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1173393</td>
<td>1764739</td>
<td>642870</td>
<td>639557</td>
<td>113663</td>
<td>368648</td>
</tr>
<tr>
<td>M</td>
<td>1515003</td>
<td>1131618</td>
<td>339595</td>
<td>638609</td>
<td>116970</td>
<td>36445</td>
</tr>
<tr>
<td>O</td>
<td>1615715</td>
<td>101260</td>
<td>560419</td>
<td>290907</td>
<td>4155</td>
<td>158245</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3524147</td>
<td>1504284</td>
<td>1568236</td>
<td>234789</td>
<td>563337</td>
<td>1419041</td>
</tr>
</tbody>
</table>

B3: S—single-family, M—multi-residential, O—other.

G2: G1—EH mainly, G3—EH partly.

H6: electric radiators, GCHP—ground-source heat pumps, EAHP—exhaust air-source heat pumps, ASHP—air-source heat pumps.

References


[18] Yin RK. Case study research. SAGE; 2009.


