INTERNATIONAL BUILDING PERFORMANCE SIMULATION ASSOCIATION ENGLAND

Proceedings of BSO-V 2020
5th Building Simulation and Optimization Virtual Conference
21-22 September 2020
Loughborough, UK

An urban-scale residential stock model for grid-constrained regions

Peter McCallum[1], David P. Jenkins[1], Paraskevi Vatougiou[1]
[1]Urban Energy Research Group, Heriot-Watt University, UK

Abstract

This work was conducted to enhance residential demand modelling for energy system integration in grid-constrained regions. The Orkney Islands are used in this study to contextualise new methods for physics-integrated thermal stock models, using data from the UK's Energy Performance Certificate (EPC) register, the Home Analytics database, and an open GIS database.

A localised urban district is studied in detail (population ~500) in which an automated archetyping process captures key geometry features of the constituent buildings through a parametric dwelling model, whilst integrating the various underlying datasets. To derive thermal demands, the model compiles and runs large numbers of EnergyPlus simulations, to give transient demands for a full year. Preliminary model outputs are discussed; however, the focus of this work is the encompassing framework and novel parametric dwelling model, which is central to the archetyping scheme. Discussions are provided on obstacles to scaling the model to much larger regions; concerns over manual interventions during input verification exercises are considerably more likely to limit scale, than computational concerns. This newly developed package – 'ParaDwell.jl' – was written in Julia, and is publicly available on GitHub.

Introduction

The recent commitment in the UK to phase out new installations of fossil fuel based heating systems has left households with uncertainty as to what alternatives offer the most cost effective heating solutions, both over medium and long term. Distribution System Operators (DSOs) are indirectly subjected to this uncertainty, and have an ever growing need to protect and plan networks to meet the imminent transformation of domestic loads. Coupled with substantial renewable supply contributions from wind and solar, it is now imperative that we reduce uncertainty in our estimations of temporally refined energy demands.

Urban Building Energy Modelling (UBEM) tools have demonstrated significant potential in energy network planning. Sola et al. (2020) identified a set of major UBEM tools, underpinned by thermo-physical Building Energy Performance Simulations (BEPS) including EnergyPlus, ESP-r, and Modelica. The highlighted models with moderate to high temporal resolution ($\leqslant$1 hour timestep) included: CHREM (Swan et al., 2009); BEM-TEB (Bueno et al., 2012); umi (Reinhart et al., 2013); TEASER (Lauster et al., 2016; Remmen et al., 2016, 2018); CityBES (Chen et al., 2017). Of these, umi, TEASER, and CityBES are configured to interpret GIS data when generating building stock, the latter two using the open CityGML model. Others have used the CitySim+ thermal engine in conjunction with CityGML (Rosser et al., 2019).

Whilst these are very powerful tools, access and availability of CityGML data can inhibit widespread application both within urban and rural settings. As a variation of this approach, Schiefelbein et al. (2019) applied TEASER alongside OpenStreetMap (OSM), modelling 55 houses using a reduced order model. Along with key geometrical features, building properties were assigned automatically in TEASER using an archetype approach, based on stock characteristics and basic building input data.

A potentially richer description of building characteristics can be accessed through dedicated building energy efficiency survey databases – Energy Performance Certificate (EPC) registers, as implemented within the European Union in accordance with the Energy Performance of Buildings Directive (EPBD). Depending on the extent of EPC records in a target region, this can, in theory, provide lower uncertainties with respect to building fabric classification, degree of refurbishment, building system information and occurrence of extension work (with respective construction period), particularly when EPC records can be identified between small clusters of neighbouring houses. These surveys and databases also come with uncertainty; however, the subject of quantifying this uncertainty is well established (Corrado and Mechri, 2009), allowing a tangible goal of quantifying one facet of modelling uncertainty, given an understanding of input uncertainty.

EPCs were used for this purpose by Ali et al. (2019), who demonstrated a data-driven approach to archetype formation, finding significant disparities between analyses using archetypes generated at national and local-level using Irish building stock. Pasichnyi et al. (2019) also used EPCs via an archetype approach, accessing both EPC data and hourly-metered heat demand data (all fed by one of Stockholm's expansive district heating networks). In contrast to the other themes already highlighted, archetype building models were manually created in DesignBuilder (with EnergyPlus BEPS engine), using three archetypes to describe a relatively homogeneous section of Swedish building stock.

Finally, Jorissen et al. (2018) provided the powerful open source Modelica library – Integrated District Energy Assessment by Simulation (IDEAS). This falls within the simulation framework, OpenIDEAS, which includes GIS-integration developments via TEASER (De Jaeger et al., 2018). Furthermore, the Python module StROBe (Baetens and Saelens, 2016) provides various forms of stochastic residential occupancy behaviour.

In spite of these examples, where stochastic patterns and metered data were used to configure the temporal response of models, open questions still exist around the integration of GIS, building survey, and temporal control data; each having inherently different taxonomy. Furthermore, in most applications there is an underlying need for large numbers of automatically generated context-specific archetypes, exploiting building forms that are mapped out by GIS databases. The present work demonstrates application of a model which approaches integration of the three themes highlighted above:

1. automated building form archetyping using GIS;
2. use of EPCs for building characteristics and properties;
3. data-driven dynamic thermal control patterns for stochastic representation of household behaviour, using smart meter data.

The presented work is based around preliminary application of the ParaDwell.jl model, specifically in relation to points 1 and 2 above. The model is used here to study a semi-urban district in the Orkney Islands, a remote archipelago in the far North of the UK. Despite this contrast to much of the dense-urban case work in existing literature, Orkney presents a series of relevant challenges which test general application and replicability to different topologies, and data availability constraints. It also incorporates a number of key features of an advanced energy system: substantial renewable generation capacity (including wind, solar and marine); active network management; hydrogen production and consumption; high rate of electric vehicle uptake (within UK context); a large proportion of electrified heating systems (with increasing use of heat pumps); and plans for commercialised aggregation services through distributed domestic battery storage. Furthermore, Orkney became one of the first regions of the UK to fully experience the effects of grid constraint issues, due to a combination of prohibitive sub-sea interconnect cost (and planning) issues, and sustained annual net export balances dating back to 2013, as a result of high levels of renewable generation and modest demand.

Simulation

The model build process within ParaDwell.jl (Figure 1) is largely dictated by the three main classes of data source: GIS data for building form descriptions; survey and analytics databases (EPC and Home Analytics (HA), in the UK context) for stock characteristics and building properties; and smart meter data for temporal aspects of behaviour (this is the subject of ongoing work).

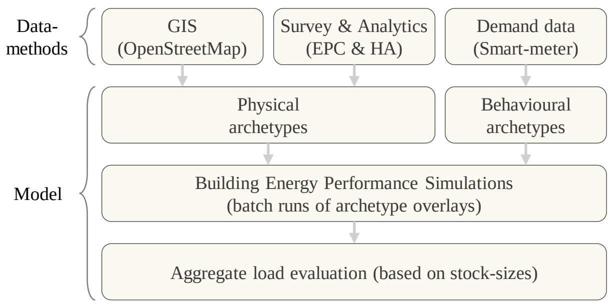

Figure 1: Model build process for ParaDwell.jl.

Whilst the first two data classes involve distinct activities during construction of the model, they do become fully integrated through a series of steps which ultimately results in a set of *physical archetypes*. The third data class, however, is fully distinct both in terms of data process and archetypal nature. These temporal activity constraints define household practices rather than building properties, so can be considered *behavioural archetypes*. In order to simulate the stock, both archetype schemes are combined, resulting in a large number of simplified BEPSs. A final aggregation stage is then carried out by weighting the simulation outputs for the various overlaid archetypes.

Before detailing the processes involved with respect to the data-methods, the central parametric dwelling model is described, which forms the foundation of the physical archetype.

Parametric dwelling model

Purpose and design

The parametric dwelling model is designed to automate a process which captures the shape and basic properties of a dwelling using a minimal set of parameters. It is used to relay information from the various data sources, which is inherently contextual, to a parametric abstraction that can be archetyped, and dispatched to an automated `.idf` file generator.

Principal dimensions, construction type/age, number of storeys, adjacencies, and heating systems are all assessed before being incorporated into an archetype label (a large string). This system is designed in a way to allow suitably similar dwellings to be assigned the same label, by observing a series of analytical tests around geometry and stock data records. Data which is unique to each individual dwelling (Unique Property Reference Number (UPRN) and grid-references or latitudes/longitudes) are purposefully excluded from the archetype label to allow coincidences in archetype (hence to facilitate an archetype approach), although these are still retained elsewhere in the model.

Interdependency between parameters is also observed, to ensure that there are no superfluous fields containing derived quantities (such as area), in order to keep the label unambiguous.

Archetype naming

Both archetype naming convention and `.idf` template system are designed to be easily modified for added complexity. In the scheme presented here there are 12 parameters. Six parameters are dimensional: plan width (`w`), depth (`d`), dwelling height (`h`), T-shape projection (`tp`), T-shape offset-left (`tol`), T-shape offset-right (`tor`). For basic rectangular plans, the latter three parameters are zero. Non-zero values can be used to represent T-, L-, or C-shapes. The remaining parameters describe: number of storeys within dwelling (`s`), the roof pitch angle (`rpa`), boundary conditions or adjacencies (`bc`), construction period (`age`), construction type (`con`), and heating system (`htg`).

To provide an example from the present case study, the following archetype was identified within the stock:

```
"w060-d080-h060-tp000-tol000-tor000-...
s2-rpa20-bcEEEAEG-ageHI-conTMB-htgEL(ST)"
```

This represents:

→ a 96m^2 house (8×6m internal plan, rectangular);

→ split over two storeys, 6m to eaves;

→ the roof is pitched (20°);

→ the building was constructed in SAP age bands H or I (1992-2002);

→ construction is timber; and

→ heating is via electric storage heaters (with reference to Table 1)

From the adjacencies, it is also known that this house is either end-terrace or semi-detached (which are considered the same in the BEPS model), the six character code being read in the order front, back, left, right, top, bottom; each assigned as external (`E`), adiabatic (`A`) or ground (`G`). This short description of adjacencies is able uniquely map to 18 different conceivable dwelling configurations.

Fuel	`htg`	**Heating system**
N/A	NONE	No heating
Gas	NGB(AB)	Nat. gas boiler (A-B)
	NGB(CDE)	Nat. gas boiler (C-E)
	NGB(FG)	Nat. gas boiler (F-G)
	NGF	Nat. gas fire
	LPGB(AB)	LPG boiler (A-B)
	LPGB(CDE)	LPG boiler (C-E)
	LPGB(FG)	LPG boiler (F-G)
	LPGF	LPG fire
Oil	OILB(AB)	Oil boiler (A-B)
	OILB(CDE)	Oil boiler (C-E)
	OILB(FG)	Oil boiler (F-G)
	OIL(RNG)	Oil range
Elec.	EL(RES)	Electric resistive boiler
	EL(ST)	Electric storage heating
	EL(RH)	Electric radiant heating
	EL(UF)	Electric underfloor heating
	ASHP(AW)	ASHP (air to water)
	ASHP(AA)	ASHP (air to air)
	GSHP(GL)	GSHP (ground-loop)
	GSHP(BH)	GSHP (borehole)
Solid	SLD(OF)	Open fire
	SLD(WBS)	Wood burning stove
	SLD(BIO)	Biomass boiler
	SLD(RNG)	Range

Table 1: Reference table for `htg` parameter.

The archetype code can potentially describe hundreds of thousands of variations in building stock, predicated on the actual existence of certain real combinations. Both underlying scheme and software implementation are designed to be flexible and extensible. Prioritised extensions include building orientation (`o`), sub-floor type (`f`), and percent glazing (`pg`). With respect to their omission in the current scheme, orientation can be interpreted from the GIS data; however, it has been avoided in the present case to simplify the tests. Suspended or solid floor constructions are not specified in the HA data; therefore, more advanced data methods are required extrapolate EPC data to all dwellings. Generalised methods for glazing characterisation is a notorious issue in UBEM application; present assumptions in this model are the subject of ongoing review in light of new methods and data.

Usage: data-methods

The parametric building model is called during two distinct steps between assessing GIS data, and linking this to Energy Performance Certificate (EPC) and Home Analytics (HA) data.

The model is used initially to assess *building* footprints obtained from OpenStreetMap (OSM) features, which can contain multiple dwellings. The geometric parameters are defined at this stage (wherever possible), determining six principal dimension (from which the basic rectangular or T/L/C-shaped plans can be inferred). This is described in a later section.

Building vertex grid references, in terms of their representation within the GIS data, are approximations, as are the dimensions evaluated here. Retaining excessive precision becomes arbitrary (i.e. representing building dimensions in millimetres), given the unquantified uncertainty in the GIS data. Moreover, in the interest of rationalising dwellings into archetypes, a rounding factor (l_{arch}) is applied. For this proof-of-concept study, $l_{arch} = 2$m was used (in general application, this would be smaller). For the purpose of generating aggregate demands for large building stock, this approach is considered appropriate, whereby the distribution of building shapes and sizes become grouped into representative averages. The parameter l_{arch} is also amenable to verification studies.

A second stage is then carried out on the subdivided buildings, where dimensions are corrected to reflect the *individual dwellings* (including area checks against survey and analytics databases), adjacencies are established based on dwelling type, and the remaining parameters are are populated with dwelling construction/system data records.

Usage: BEPS model

Once a specific stock has been assessed, the list of archetype labels can be separated into their parameters and used to generate the EnergyPlus simulation input files (`.idf`). A series of steps are then followed to generate all archetype `.idf` files without the need for any user intervention.

1. The model contains a library of template `idf` files, each representing a different heating system type; the model selects the appropriate template based on the `htg` parameter.
2. A further library of `.idf`-snippets is then used to introduce the generic building form, in terms of parametric shape (conditional on zero/sign of `tol,tor,tp`) and adjacencies (`bc`).
3. Schedules are introduced, generated through a stochastic process using representative smart meter data (outside the present scope). Other tools, such as StROBe, are also being explored.
4. Sections/lines of code are expanded to represent multi-zone system components where required (currently applies to non-rectangular plans and dwellings of two or more storeys; in future this will also include room-specific zones).
5. The model is populated with all remaining detail from the parameters, including vertex coordinates for all building surfaces, calculated window vertices (based on assumed glazing proportions of 25% wall area), roof pitch configuration (as per `rpa`), and materials assignments in accordance with the `age` and `con` parameters.

GIS-linked archetyping method

The following outlines the data methods undertaken on the GIS data, to allow archetyping at subsequent stages:

1. OpenStreetMap (OSM) data is accessed and collected for the region of interest
2. Particular attention is directed towards building features; however, further use is also made of limited additional features to aid contextual studies when mapping, including roads and coastlines.
3. To interpret all available building forms, the coordinates of each 2D vertex are interpreted from the relevant nodes in the corresponding XML object, before taking further steps to characterise each building.
4. Building boundary lengths and orientations are assessed using formal geodesy methods (the mathematical field for non-spherical, global geometry definition).
5. Where possible, the building forms are characterised into simple shape groups. An analytical procedure is then followed to interpret the dimensional parameters:
   (a) building boundary orientations are normalised to ensure each consecutive boundary meets at exactly 90°, providing that the original orientations give an approximate orthogonal plan;
   (b) the number of vertices is observed: four, six and eight vertices can be assessed as rectangular, L-shaped, and T/C-shaped, respectively;
   (c) the sequence of connecting angles is followed to determine shape, before assigning the principal dimensions to the archetype label.

   Where exceptions occur with respect to non-orthogonality or vertex count, the buildings are considered unclassified. At present, these instances are approximated using rectangular plans of representative size and aspect ratio. Future plans include a bespoke (unique) building model for each identified case.

What results from this stage is an intermediary set of data in the form of simple 2D geometric building footprints, which is worth noting, has limited relevance outside this present analytical-routine. In order to produce a set of *'physical archetypes'* (as Figure 1) which can generically be used to characterise building stock, theEnergy Performance Certificate (EPC) and Home Analytics (HA) data must be used to determine how many dwellings are within each building footprint, before including further information on storeys, constructions and systems.

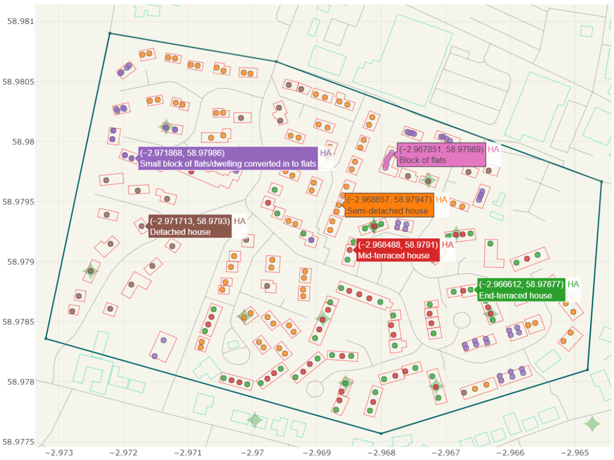

Figure 2: Mapped Home Analytics data, showing dwelling types.

Figure 3: Mapped EPC data, showing age-bands (key removed for anonymity).

EPC-linked stock characterisation

Linking to OpenStreetMap data

Vertex coordinates of buildings which are obtained from OpenStreetMap (OSM) features are retained within the model during the build process to link GIS data with building stock survey and analytics databases. Dwelling grid references provided in HA database are mapped to OSM data by using a geometrical routine to find points that lie within each feature. The OSM features become considered residential buildings if it is now associated with one or more Unique Property Reference Number (UPRN). If a feature has no UPRN it is excluded from analysis, as a non-residential building or other structure (free-standing garage, substation etc.). The HA data also includes a unique 'Building Block ID', which assists when identifying multi-dwelling buildings. Having assigned UPRNs to specific building footprints, EPC data can now be associated with these objects.

Figures 2 and 3 show the mapping of selected parameters which are used to populate the dwelling model.

Predicting building storeys

An approximate process was adopted for building storey classification. From knowledge of UPRNs which lie within each building footprint, the sum of the recorded dwelling areas, $\sum^{block} A_{epc,ha}$ (EPC data taking precedence over HA where available), can be divided by the GIS-measured area, A_{gis}, to infer the number of building storeys, acknowledging that $A_{epc,ha}$ and A_{gis} relate to internal and external areas, respectively. For the stock within the case study, the distribution of this ratio was considered, see Figure 4. Around an empirical cut-off value of $f_s = 1.0$, 96.2% of shape-categorised building agreed from inspection with:

$$\mathtt{s} = \left\lceil \frac{\sum^{block} A_{epc,ha}/A_{gis}}{f_s} \right\rceil \tag{1}$$

where $\mathtt{s}$ is the building storeys parameter. Buildings for which Eqn.(1) failed featured errors in the HA area (with respect to neighbouring EPCs), errors in A_{gis}, or included features such as integral garages.

Dimensions and adjacencies

For detached dwellings the building footprint becomes the dwelling footprint once re-scaled using the ratio $A_{epc,ha}/A_{gis}$ (with archetypes now described in terms of internal area). For multi-dwelling buildings, rectangular plans are split according to dwelling type: semi-detached and terraced houses (based on HA data) involve subdividing the building footprint, with adjacencies assigned accordingly. Where flats are identified, buildings are split on the basis of patterns that would be expected for a block of a certain number of storeys, although there is no current method to determine where a particular UPRN is placed within that sequence. All multi-dwelling buildings which are not rectangular are converted to a representative rectangular plan, which respects plan depth and area, prior to subdivision. A scaling factor of $\sum^{block} A_{epc,ha}/A_{gis}$ is used for all multi-dwelling buildings.

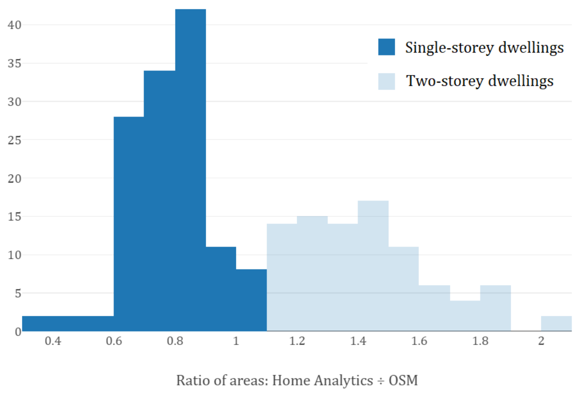

Figure 4: Distribution of $\sum^{block} A_{epc,ha}/A_{gis}$

Age-bands, constructions, and heating systems

The remaining parameter fields are populated with EPC data. Where EPC data is not available, HA data is used. Each piece of data is mapped to a rationalised list of options; in the case of age-bands, the SAP labels A-L are used; however, these area grouped in the same manner as the HA database:

- A (pre-1919)
- B-C (1919-1949)
- D-F (1950-1983)
- G (1984-1991)
- H-I (1992-2002)
- J-L (post-2002)

Constructions can be: timber; cavity wall; brick/stone; or system-built. Heating systems are labelled with a short code based on Table 1, allowing reference to gas, electric, HP, oil, or solid. Other future additions, such as district or hydrogen may be added, along with expanded descriptions of heating efficiencies and sub-types.

Results and Discussion

The various stages through data-methods, archetyping, simulation and aggregation, as depicted in Figure 1, present different issues. Building and running the simulation (third row of Figure 1) is straightforward once a robust procedure is developed for writing input files and executing the BEPS engine (EnergyPlus). Processing an aggregate demand (fourth row) is very straightforward; however, validation activities start at this stage, initiating a crucial and time-consuming task, invariably requiring model reconfiguration and more detailed input verification.

The biggest challenges lie in accessing, processing, verifying and cross-linking the necessary datasets (first and second rows, Figure 1), both when configuring the first runs, and reconfiguring during validation. These issues are discussed below.

Data-methods

Within the three classes of data used for building the model, it is generally best to access as many high-quality sources as possible. For automation purposes, it is essential to have some mechanism to cross-reference the different sets, which can be achieve using intermediary datasets if required. In this work, for example, it was not possible to directly link the GIS (OpenSteetMap) data directly to the Energy Performance Certificate (EPC) database, as grid-references were not provided in the latter. Home Analytics (HA) data , however, could be related to both GIS data and EPC data via grid-references and Unique Property Reference Number (UPRN), respectively, allowing an indirect link between GIS and EPC. Any further databases that can provide valuable building stock context which also references UPRN, could be introduced in the same manner.

The value in accessing additional high-quality data wherever available is to support model-input verification. Furthermore, by observing the fundamental distinction between model verification and input verification, appropriate measures can be developed that best address the gap between simulated results and validation data.

Regarding data input verification of the present model, HA data was used as basis data, as it covers almost all dwellings. 259 UPRNs were recovered, giving input data for area, dwelling type, construction type, age-band, and heating system. 109 EPCs were then identified with matching UPRNs. Relevant conflicts between EPCs and HA were examined, and depending on field, the most appropriate data took precedence. Age, heating system, construction, and area were all taken from EPCs where available. Dwelling type, however, was more consistent in HA data.
As an integral part of the ParaDwell.jl model, the mapping interface was developed to simultaneously overlay GIS, EPC, and HA data. This single feature is essential to the data validation process, and allowed efficient spot checking of gaps and inconsistencies. Clear errors could be identified, for example, where adjoining dwellings were labelled with different construction eras; in such instances, it was possible to override errors in the EPC data with HA data. The present work was limited to analytical and conditional routines, with user observation; however, future work aims to automate much of this process using machine learning methods.

Archetyping

The parametric model was central to the archetyping process. Difficulties generally surrounded the categorisation of complex footprint shapes; there is also no current method to subdivide any shapes other than rectangles, such instances were replaced with a suitable rectangular plan.

Another major difficulty was the characterisation of building heights. The current process performed reasonably well but featured persistent issues for various reasons. Errors in area records/evaluations caused some issues. For more general use, integral garages, loft conversions, basements, split levels, and mixed-height terraces will all present issues.

Other significant areas for development include methods for determining glazing proportions and characterisation of roof pitches, dormers etc., which are likely to require considerably richer data input.

The next major update to the model will be the inclusion of a new module for Ordnance Survey (OS) data interpretation. Preliminary work using OS data has demonstrated significant advances with regard to the above issues. In particular:

- multi-dwelling buildings (such as terraces) are generally subdivided, making archetype definitions significantly easier, especially for more complex footprints;
- the OS Building Height Attribute (BHA) allows definitive interpretation of the number of storeys, addressing the single biggest factor in manual error handling;
- BHA also defines apex and eaves heights, helping infer roof pitch and identify flat roofs;
- dwellings in the OS data can include independent features such as garages and outbuildings, which again, improves archetype definition.

Regarding the most onerous cause of intervention in the present work (predicting the number of storeys using OSM data), examples of anomalous cases that caused skewed archetype dimensions, included:

- a single storey semi-detached dwelling adjoining a two storey semi-detached dwelling;
- a pair of flats adjoining a two storey semi-detached dwelling;
- a purpose built block of four flats, with a large open atrium separating the building into two.

In all of these cases, the respective buildings were represented as a single polygon in the OSM data. Use of OS data would make the approximation in Eq.(1) redundant, and would drastically improve the effectiveness of the model for larger regions.

The majority of the development work required for OS data integration relates to more advanced computational methods for visualising the more detailed data, along with the added requirement for 3D viewing. Ongoing work includes the introduction of GPU-based methods.

Simulation and aggregation

ParaDwell.jl provided a fully automated process for `.idf` input file generation, batch processing, result extraction, and aggregation, which is set up to run for any generic stock. The model responded well to tests from both the case study region, and other similar areas in Kirkwall. A variety of dwelling configurations were identified, with an array of different heating systems and construction types. In each case, the `.idf` generation process captured the parametric description of each archetype.

For simulation, each archetype was sequenced through the EnergyPlus command-line interface, internally within Julia (the high-level programming language used to build the model). Using a laptop with an i5 (7th gen.) processor, with 16GB RAM, simulations for the very basic (one zone per floor) archetypes took ~5 seconds. 134 archetypes were required for case study, all unique `.idf`s were generated in $\ll$1 second, and simulated in ~10 minutes.

Conclusion

The framework and model presented in this work has been developed to simulate aggregate thermal demands on an urban scale, using automated processes for integrating GIS data, dwelling survey, and analytics databases. Testing of the model has demonstrated effective application of the automation and analytic processes, in particular around the parametric dwelling model. In terms of BEPS integration, batch simulation, and aggregation, the model also performed well.

The tests also highlighted some major challenges. Considerable effort was required to manually verify input data for the relatively small building stock. Limitations also persist around complex building shapes, characterisation of glazing extent, and building heights.

In general with regard to these challenges, the rapidly advancing data landscape (in particular open data) will begin to address some shortcomings. Ordnance Survey has released an open-access version of their MasterMap, and recently extended the 3D Building Height Attribute to cover Orkney (now with almost complete coverage of the UK). The introduction of more advance analytics and data science could allow for additional layers of automation with respect to input data verification. More rudimentary analytical and empirical checks are also planned.

An objective of the present work was to study specific challenges around thermal modelling of a community with 200-300 dwellings, and to identify obstacles for scaling further. From the findings, it would be possible to expand the current approach to communities slightly larger than the present case study, as the automated processes for data interpretation were reasonably effective. However, in the absence of more advanced input verification processes, scale-limitations are likely to be due to manual interaction when (re)configuring the model. The ongoing major updates to the model for integration of Ordnance Survey data will significantly enhance scalability, by addressing a number of these input verification issues, and significantly reducing the associated manual intervention.

A further objective of this work was to design an extensible modelling framework, that can be easily augmented to add complexity, in terms of archetypes and individual Building Energy Performance Simulation models, without fundamentally re-writing the underlying code. In this regard, the model performed very well; these aspect should not present any particular issues with regard to future use at larger scales.

Regarding computational tractability, larger case studies will be necessary to determine the degree to which the stock can be rationalised in the presented archetyping scheme. The present case study featured

134 archetypes to represent 259 dwelling; at much larger scales this ratio would severely undermine scalability. However, increasing commonality in building stock would be expected for larger regions, and with the ability to simulate >10,000 archetypes in 24hrs on a standalone computer, it is unlikely that processing will fundamentally limit scale.

Acknowledgement

This work was conducted as part of the ReFELX Project (a PFER demonstrator, main funder: Innovate UK, grant: 104780), and within the scope of the National Centre for Energy Systems Integration (CESI, main funder: EPSRC, grant: EP/P001173/1).

Nomenclature

$A_{epc,ha}$	Area (m^2), from EPC database (or Home Analytics if required)
A_{gis}	Evaluated footprint area from GIS
f_s	Survey/GIS area ratio cut-off(empirical)
l_{arch}	Geometric rounding length for archetyping

References

AECOM (2018). Energy Demand Research Project: Early Smart Meter Trials, 2007-2010. *UK Data Service SN: 7591.*

Ali, U., M. H. Shamsi, C. Hoare, E. Mangina, and J. O'Donnell (2019). A data-driven approach for multi-scale building archetypes development. *Energy and Buildings 202*, 109364.

Baetens, R. and D. Saelens (2016). Modelling uncertainty in district energy simulations by stochastic residential occupant behaviour. *Journal of Building Performance Simulation 9*(4), 431–447.

Bueno, B., G. Pigeon, L. K. Norford, K. Zibouche, and C. Marchadier (2012). Development and evaluation of a building energy model integrated in the TEB scheme. *Geoscientific Model Development 5*(2), 433–448.

Chen, Y., T. Hong, and M. A. Piette (2017). Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. *Applied Energy 205*(August), 323–335.

Corrado, V. and H. E. Mechri (2009). Uncertainty and Sensitivity Analysis for Building Energy Rating. *Journal of Building Physics 33*(2), 125–156.

De Jaeger, I., G. Reynders, Y. Ma, and D. Saelens (2018). Impact of building geometry description within district energy simulations. *Energy 158*, 1060–1069.

Jorissen, F., G. Reynders, R. Baetens, D. Picard, D. Saelens, and L. Helsen (2018). Implementation and verification of the ideas building energy simulation library. *Journal of Building Performance Simulation 11*(6), 669–688.

Lauster, M., M. Mans, P. Remmen, M. Fuchs, and D. Müller (2016). Scalable Design-Driven Parameterization of Reduced Order Models Using Archetype Buildings With Teaser. *Proceedings of the 6th German-Austrian IBPSA conference* (September), 535–541.

Pasichnyi, O., J. Wallin, and O. Kordas (2019). Data-driven building archetypes for urban building energy modelling. *Energy 181*, 360–377.

Reinhart, C., T. Dogan, A. Jakubiec, T. Rakha, and A. Sang (2013). Umi - an urban simulation environment for building energy use, daylighting and walkability. *13th IBPSA Conference, Chambéry, France, August 26-28*, 476–483.

Remmen, P., M. Lauster, M. Mans, M. Fuchs, T. Osterhage, and D. Müller (2018). TEASER: an open tool for urban energy modelling of building stocks. *Journal of Building Performance Simulation 11*(1), 84–98.

Remmen, P., M. Lauster, M. Mans, T. Osterhage, and D. Müller (2016). CityGML Import And Export For Dynamic Building Performance Simulation In Modelica. *Building Simulation and Optimization conference (BSO16)*, 329–336.

Rosser, J. F., G. Long, S. Zakhary, D. S. Boyd, Y. Mao, and D. Robinson (2019). Modelling urban housing stocks for building energy simulation using CityGML energyade. *ISPRS International Journal of Geo-Information 8*, 163.

Schiefelbein, J., J. Rudnick, A. Scholl, P. Remmen, M. Fuchs, and D. Müller (2019). Automated urban energy system modeling and thermal building simulation based on OpenStreetMap data sets. *Building and Environment 149*, 630–639.

Sola, A., C. Corchero, J. Salom, and M. Sanmarti (2020). Multi-domain urban-scale energy modelling tools: A review. *Sustainable Cities and Society 54*(February 2019), 101872.

Swan, L., V. I. Ugursal, and I. Beasuoleil-Morrison (2009). Implementation of a Canadian residential energy end-use model for assessing new technology impacts. *11th International IBPSA Conference - BS2009*.