# A TAXONOMY OF KEY UNCERTAINTIES USING HIGH-LEVEL FRAMEWORKS FOR ENERGY MODELLING

*Javier Urquizo[a,d], Carlos Calderón[b], and Philip James[c]*
*[a] Newcastle University, Newcastle upon Tyne, United Kingdom, j.urquizo@newcastle.ac.uk*
*[b] Newcastle University, Newcastle upon Tyne, United Kingdom, carlos.calderon@newcastle.ac.uk*
*[c] Newcastle University, Newcastle upon Tyne, United Kingdom, philip.james@newcastle.ac.uk*
*[d] Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador, jurquizo@espol.edu.ec*

*Abstract– This paper provides a tri-dimensional taxonomy of uncertainty of the Newcastle CarbonRoute Framework (NCRF) using a concept map. It requires the identification of the sources, issues and sub-issues of the uncertainties in the modelling process. These issues can be broken down in the contributing forms of uncertainty and classified as either contributing to inaccuracy (systematic bias of the data) or imprecision (random variability of the data). Much of the data used in this research comes from surveys based on samples; some inaccuracy is unavoidable in the energy estimations presented. The most significant source of inaccuracy is perhaps the sampling error, where the characteristics of a sample do not exactly match the characteristics of the whole population. The purpose of the research is to develop a taxonomy that shows how uncertainties are propagated through the modelling process (data – model – refinement – validation) and in the resulting estimates of annual energy consumption.*

## I. A TAXONOMY OF KEY UNCERTAINTIES USING HIGH-LEVEL FRAMEWORKS

This research inherited the Newcastle Carbon Route Map (NCRM)(Calderón *et al.*, 2012), which is an early incarnation of a building level data set for Newcastle upon Tyne, UK. The initial phase of this research involved substantial data management, cleaning, restructuring and additions to this initial data set. The resultant data set incorporated in a single database table a large number of building related data sets. The Newcastle CarbonRoute Map Framework (NCRF) utilises this data set and adds on the energy modelling aspect through linking with the English House Survey (EHS) as input to the Cambridge Housing Model (CHM) (Calderón *et al.*, 2015). This provides the means to produce building level energy consumption estimates which in turn can be analysed both spatially (Urquizo *et al.*, 2018a) and aspatially (e.g. by building type). This building level approach through the NCRF provides the potential for energy planners and other bodies to model energy interventions with flexibility in scale and to potentially adapt plans to local area characteristics (Urquizo *et al.*, 2016; Urquizo *et al.*, 2017).

This section proposes a classification of the key uncertainties in the energy model. To have an understanding of the uncertainties of the NCRF outcome, Fig.1 shows a three dimension integrative taxonomy of the uncertainty adapted form (Han et al., 2011, p. 7) by identifying the nature (location), the cause (level) and the extent (nature) of the uncertainty

i. The first dimension is the location (source) dimension of uncertainty related to where the uncertainty manifests within the complex energy model;

ii. The second dimension is the level (issues) dimension of uncertainty related to where the substantive issues (and from there the sub-issues if meaningful) of uncertainty manifests along the spectrum between deterministic knowledge and total ignorance;

iii. The third dimension is the nature (locus[1]) dimension of uncertainty which relates to whether the uncertainty is due to the lack of knowledge or is due to the inherent variability of the variable being described.

The model outcome uncertainty in Fig. 1 is the accumulated uncertainty caused by the uncertainties in all of the locations (context, model, inputs to the energy model, and refinement) that are propagated through and are reflected in the resulting estimates of annual energy consumption (aggregated in geographic boundaries or repeated for the same property type, area and year of construction). This uncertainty outcome could be considered a prediction error, since it is the discrepancy with the UK Department of Energy and Climate Change (DECC) median value.

Fig. 1 accounts for the aggregate of uncertainties in all sources. However, it should be noted that NCRF could be used as an energy policy analysis model and estimate energy at other boundaries where there is no aggregation (beyond DECC known values), i.e. to estimate annual energy consumption outcomes for aggregates (or repeated) situations where DECC values are not publicly available. For these cases, the taxonomy shown in Fig.1 is still valid.

In Fig.1, the uncertainty issues are related to the energy modelling methodology. The first dimension "Location" of uncertainty refers to: analytical approach, domestic energy model, full Standard Assessment Procedure (SAP) input and refinement/validation. This section explains the three dimensions associated with the "analytical approach," leading to Section 2 on the uncertainty in the other sources identified, namely the "domestic energy model", "full SAP input" and "refinement/validation."
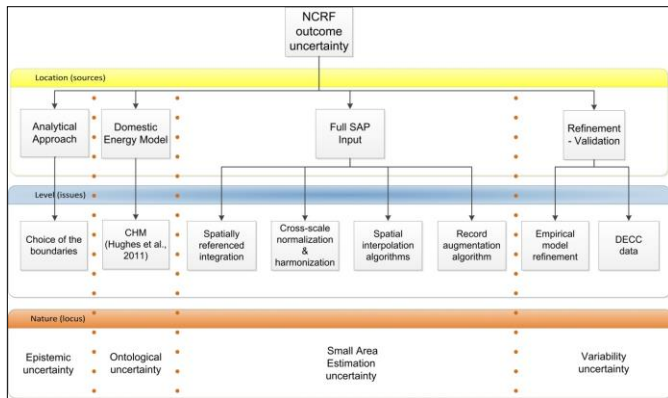
---

[1] A centre (focus) of great activity.

Fig.**1** NCRF outcome uncertainties

The analytical approach refers to the conditions and circumstances that underlie the choice of the boundaries of the system, the framing of the concepts and the terminology of the research question to be addressed within those boundaries. In this research, the term analytical approach refers to the following issues: (i) the energy model and (ii) the model complexity. The NCRF energy estimates correspond to an engineering method (EM) which calculates the energy consumption of end-uses for dwellings based on the heat transfer and thermodynamic properties. Model complexity arises from the fact that NCRF has multiple inputs at different scales. It has two data sets at a resolution of the individual dwelling, one data set of rough approximations of household occupancy and three average regional scale landscape and climatic data sets.

The third dimension of the uncertainty is the nature of uncertainty. An important feature of uncertainty is the distinction between: (i) epistemic uncertainty (the uncertainty due to the imperfection of our knowledge), which may be reduced by doing more research and using added empirical efforts; and, (ii) variability uncertainty, which is due to the inherent variability of the data. Between these two extremes, there is "ontological uncertainty", which can be seen as having a semi-structured uncertainty, and "small area estimation uncertainty," which can be seen as having a semi-variability uncertainty, as shown in columns of Fig.1.

The arrows in Fig.1 associate the NCRF outcome uncertainty with the first dimension sources, and in turn associate each sources with its issues. The figure also presents the outer left source analytical approach having a structural (epistemic uncertainty) in the locus third dimension and from there an increasing parametric uncertainty at the far right with the refinement –the validation source.

CHM uses standard parameters that do not fit with the local area characteristics, then a discrepancy between the modelled and measured energy consumption can be observed. This means that this research has to consider the structural uncertainties in the taxonomy of Fig.1. Also, the "CHM model" is an idealized model of the domestic stock, and there

is the possibility of an undetected error in the design that introduces "ontological uncertainty". As an example, CHM does not consider some energy saving/generation technologies like the small-scale hydro-electric generator which is being considered in SAP 2009 (BRE, 2011, p. 82). The introduction of technologies that might be unfamiliar to the CHM model may carry a higher degree of ontological uncertainty.

The input data to CHM corresponds to a full SAP data set. This research performs indirect estimates from secondary sources (EHS, UK Census) in a city sample. The output estimation is the underlying expected value for any area given the independent variables included in the NCRF estimates and not the real value for the "small area" in question. For this reason, as this is not a direct measure of the constructed SAP record for each dwelling in the city, but rather estimation for each building, it can be considered as numeric uncertainty, i.e. towards the right hand side in Fig.1.

In summary, this section proposes a three dimensional integrative taxonomy of uncertainty representing a conceptual framework that helps to organize our knowledge by drawing our attention to relevant sources, issues and the nature of uncertainty in the NCRF estimates. Despite the importance of the domestic energy modelling in sub-city areas, the energy sector lacks a rigorous analytical framework to account for the uncertainties. The most common practice is to assign a single uncertainty value to the modelled output uncertainty.

## II. SOURCES OF PARAMETRIC UNCERTAINTY USING A CONCEPT MAP ION

This section explains the second dimension (issues) for the sources of parametric uncertainty using a concept map. The main idea is to show the key issues that connect and relate to the main sources of uncertainty and rank them with the most general, with inclusive issues coming first, and then links to smaller, more specific concepts until it reaches the quantification of the uncertainty in terms of an inaccuracy or imprecision. The Concept Map (CM) is proposed as a human friendly knowledge-representation of uncertainties, and is a tool especially defined for application in the learning process. It is easy be create, and flexible and intuitive for people to understand (Shapiro and Eckroth, 1987; Novak and Cañas, 2006; Sowa, 2006).

The formalization of a CM of the quantified parametric uncertainties in the NCRF outcome is presented in Fig. 2. This is an extension of previous work published on the uncertainties of the Cambridge Housing Model (CHM) (Hughes et al., 2013, p. 161), (see Hughes et al. (2013) for the uncertainties in this source model). This research uses the CHM uncertainty model as a starting point for an emerging spatial, area-based urban, domestic energy model of uncertainties. Fig. 2 shows the CM section derived from (Hughes et al., 2013), plus the additional CM sections describing uncertainty sources from NCRM: "Full SAP input" and "refinement/validation."
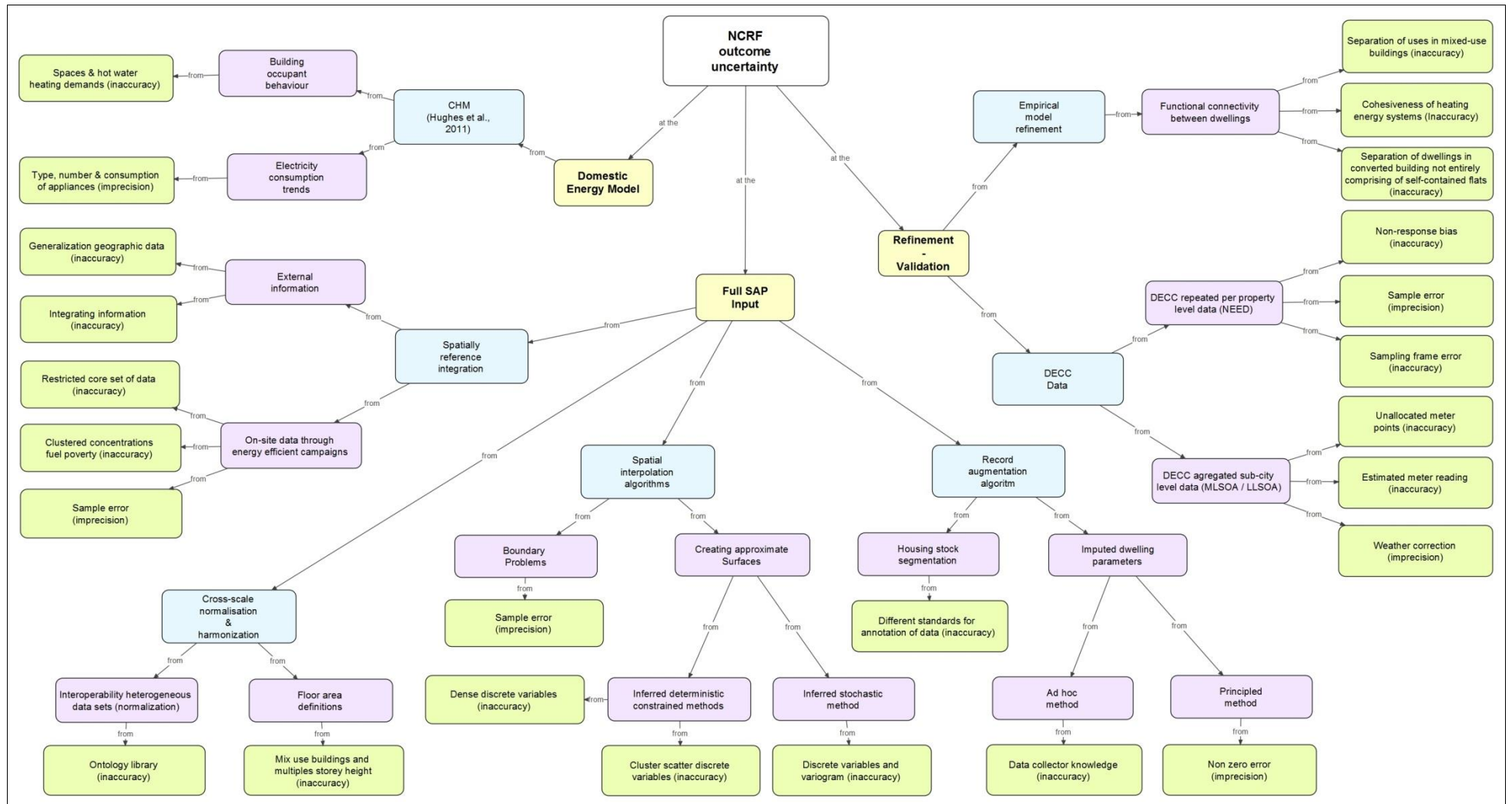
Fig. 2 NCRF outcome uncertainty concept map

In Fig. 2, the green colour corresponds to the sources of uncertainties, and orange represents the issues (activities and disagreements) causing the uncertainties in the corresponding sources. Purple represents the sub-issues (a logically visible subdivision of an issue) and finally the terminal slots in green represent a further abstraction to those quantities that cause uncertainty in terms of accuracy and precision. The accuracy is the degree of closeness of measurements (of a quantity) to that quantity's actual true value and the precision (also called the reproducibility or repeatability) is the degree to which repeated measurements under unchanged conditions show the same results.

In summary, this section extends the CHM national model through a CM representation to an area-based domestic energy model uncertainty. Hughes et al. (2013) quantified uncertainty for a national level system. Here, the quantification of spatial uncertainty, at a local level could benefit policy makers and stakeholders as it potentially legitimizes the modelling process and makes it more transparent. This Section explains the contributing forms of uncertainty of individual issues of Fig. 2.

This section develops the contributing forms of uncertainty in the "full SAP input" and "refinement/validation" sources of uncertainty. The issues are presented in anti-clockwise order (around the NCRF uncertainty outcome in Fig. 2), which also corresponds to left-right order of the issues of Fig.1.

*A. CHM Model level*

Issues related to the CHM are shown in Fig. 3. The sub-issues are the building occupants' behaviour and electricity consumption in households.

i. Building occupant behaviour.

The behaviour of occupants in a building is dependent mainly on the make-up of the household (number of people in the household, and their ages, etc.). The space and hot water heating energy consumption are largely determined by building occupant behaviour. CHM models building occupancy in a very simplistic way, which results in uncertainty in our outputs.
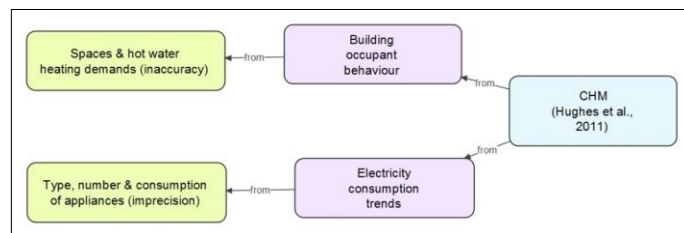


Fig. 3 Uncertainties in the CHM model level

ii. Electricity consumption in households.

Electricity consumption in domestic buildings is determined by two main factors: the type and number of electrical appliances in the property, and the use of these appliances by the occupants of the building. In houses with similar built forms, there can be a wide range of different appliances in use and these appliances presumably have a range of different power consumptions.

Occupants do influence the electricity use of a dwelling both by their purchase of electrical appliances and through their use of these appliances. The variation in the type of electrical appliances present and the occupant use of these makes domestic electricity consumption difficult to predict with accuracy. CHM does not model the appliance use in the households, which causes uncertainty.

*B. Spatial Reference Integration*

Issues in spatial reference integration follow the concept map segment in Fig. 4. The sub-issues come from the integrated data: the on-site primary survey data from energy efficiency campaigns to eradicate fuel poverty (Warm Zones) and the secondary (externally generated) data from the English Housing Condition survey.

Data sources and building characteristics are composed of both externally generated information and on-site data obtained through energy efficiency campaigns.
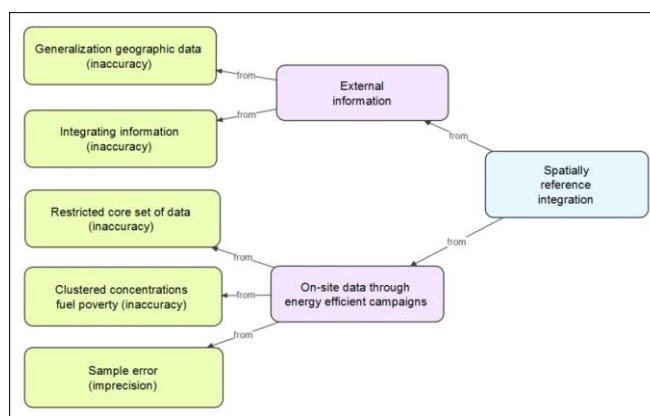


Fig. 4 Uncertainties in the spatially referenced integration

Issues from externally generated information arise in the following sub issues, including:

i. Generalization of geographic data (inaccuracy), which comes from handling different scales of data and different data formats. The building outlines (MasterMap™) are provided by the Ordnance Survey in a vector format with a unique Topographic Identification (TOID). NCRM Gazetteer, the local database, is a vector format with a Unique Property Reference Number (UPRN). In the case of a building with several apartments, a group of UPRNs corresponds to a single TOID. Census information is only available at the minimum small scale of a Super Output Area (SOA). The

thermal and the LIDAR images are raster data. The LIDAR image has a 50cm planimetric accuracy for 1km of flight altitude.

Uncertainty arises at object generalization when this research tries to combine these data, e.g. when inferring socio-economical information for a single parcel. As census information is available only at SOA, critical socio-economic information has to be considered as average over all parcels in this area. Additional generalization occurs when trying to identify in the heat loss database buildings of less than 20sqm, and it does not work at all on those less than 10sqm.

ii. Uncertainty arises at integrating spatial data from different scales, e.g. overlaying a LLSOA map over a NCRM Gazetteer map leaves some city parcels being part of several LLSOAs.

Uncertainties from on-site data are those obtained through domestic energy efficiency campaigns (Warm Zone) to alleviate fuel poverty.

The Warm Zone's approach is a door to door assessment where assessment teams systematically contacted households, mainly face to face on the doorstep to acquire information to provide (through secondary information) a measure of the fuel poverty status of the household. This information is then used to target the provision of an appropriate energy efficiency measure.

Progress against the government's target for fuel poverty reduction, as set out in the UK Fuel Poverty Strategy, is monitored using defined, detailed calculations (DECC, 2013). It is not realistic to replicate this level of data collection through a door to door assessment. Examples of uncertainty arising from the NCRM WarmZone data are:

i.  Warm Zone is restricted to a core set of data (inaccuracy), as it is largely related to the assessment of fuel poverty, both before and after intervention, and therefore is an approximate measure;

ii. Warm Zone data collection was focussed on areas where there was existing evidence of spatial clustered concentrations of fuel poverty within a locality that can benefit from high impact approach zones, which could bias the survey (inaccuracy); and

iii. Sample error (imprecision) as the sample may not be (and perhaps is not) representative of the whole MLSOA due to the purpose of the survey.

*C.  Cross-scale normalization and harmonization*

In this research, normalization is the set of standard statistical procedures used to combine two studies and cross-scale harmonization is used to combine data sets from different scales. Issues from cross-scale normalization and harmonization follow the concept map segment in Fig. 5.

i.  Semantic interoperability. The large amount of data being accumulated in NCRM needs to be adequately annotated. The variables used in exchanging and integrating information must adopt standards for the annotation of data, in order to enable consistent information retrieval, i.e. NCRM Cities Revealed and EHS age band field names need to be the same categories. Ontologies play an essential role in this integration, enabling the semantic interoperability of heterogeneous distributed systems. However, at present different building classification schemes do not use the same age bands so bands may be split proportionally to align data sets.
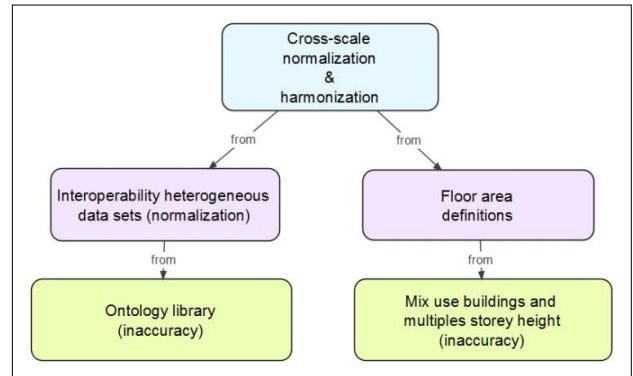


Fig. 5 Uncertainties in the cross-scale normalization and harmonization

ii. Floor area definition must be allocated for every dwelling from information available for number of storeys, height and footprint; this is estimated (inaccuracy) in mixed-use and multiple storey buildings.

*D.  Spatial interpolation algorithms*

The spatial interpolation methods are a practical way of group-based estimation. This research uses three different interpolation methods to estimate dwelling parameters' Nearest Neighbour (NN), Inverse Distance Weighting (IDW) and kriging (Urquizo *et al.*, 2018b). All methods are weighted average methods. Issues in this level are described in the concept map segment in Fig. 6.

i.  Boundary problem

Spatial interpolation technique estimates should only be used within the convex hull of the input points, and the LLSOA areas do not perfectly align with the convex hull. Areas close to the boundary may be more uncertain as the input points used in the interpolation may be restricted (imprecision).

ii. Creating approximate surfaces

This research uses two deterministic methods in the spatial interpolation algorithms: the Inverse Distance Weighting (IDW) and Nearest Neighbour (NN). These do not incorporate statistical probability theory into the development of the prediction's surfaces. Instead, these

methods use mathematical formulae (inaccuracy). IDW is a deterministic method that uses a weighted average of nearby points with distance being the only factor influencing the calculation of the weight, and NN is a deterministic method that involves identifying the closest measured point to an unmeasured point and assigning the value of the measured point to the unmeasured point.
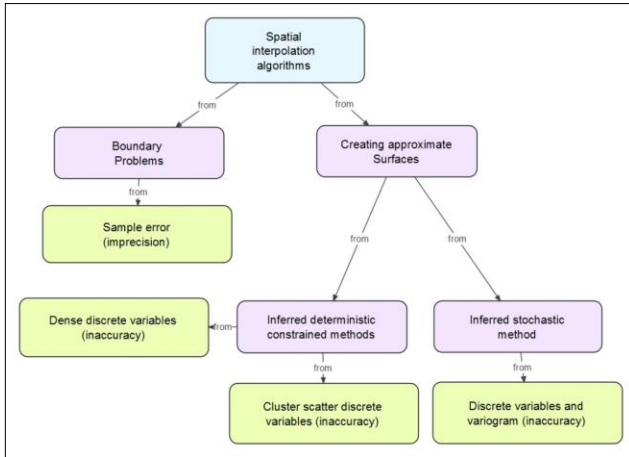


Fig. 6 Uncertainties in the spatial interpolation algorithms

Kriging instead is a stochastic method because it assigns weights based not only on the distance between surrounding points but also on the spatial autocorrelation (imprecision) between the measured points, which is determined by modelling the variability between points as a function of separation distance.

### E. Record augmentation algorithm

The record augmentation algorithm works in two stages, by first creating a physical record (age, infrastructure and land use) for grouping similar individual dwelling, and second by applying a multiple imputation procedure to find the best record within the EHS data set. Uncertainties from the record augmentation algorithm arise in the following activities:

i. Housing stock segmentation
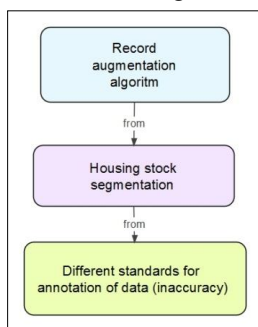Issues in the housing stock segmentation follow the concept map segment shown in Fig. 7.



Fig. 7 Uncertainties from housing stock segmentation

The issues appear in the process of augmenting information in NCRM from EHS. There is a need to adopt standards in the annotation of data, in order to consistently enable information to be imputed to NCRM from EHS in order to create a full SAP input and apply later to the CHM energy model. The main issue highlighted here is the different age bands used for property classification in NCRM Cities Revealed, EHS and NEED.

ii. Imputed dwelling parameters
Issues from imputed dwelling parameters follow the concept map segment in Fig. 8. The record augmentation schema selected is either single or multiple imputations with three possible results: an exact match, a best candidate (both in the multiple imputation schemes) and no found outcome (in a single imputation scheme). The "not found" results require an ad-hoc procedure of single imputation and the exact and best candidates use a principle component method.
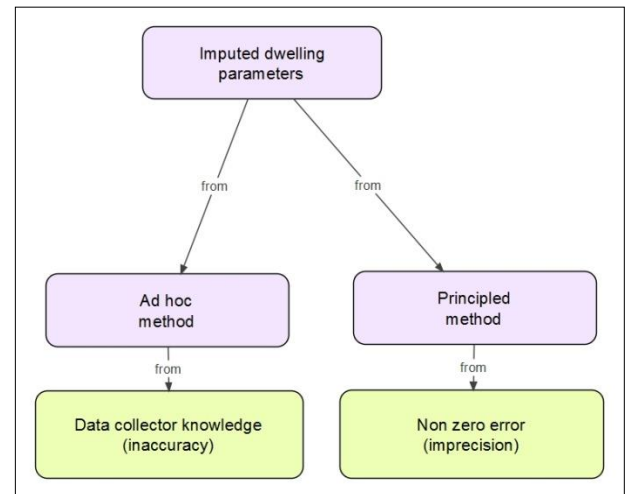


Fig. 8 Uncertainties from the imputed dwelling parameters

- The Ad Hoc Imputation method fills in missing values using the knowledge of the data collector; there is inaccuracy, which is a source of uncertainty.
- The Principle Component analysis method provides a match based on clusters of variables that are highly correlated but it will not match all variables between NCRM and EHS records; therefore, it should be considered a source of uncertainty (imprecision).

### F. Empirical model refinement

The model refinement provides an empirical approach to determine whether a dwelling belongs to one of the special cases of building properties such as: (i) Homes in Multiple Occupation; (ii) energy systems –district and group heating;

and (iii) multi-use buildings. Issues at this level are shown in the concept map segment in Fig. 9.
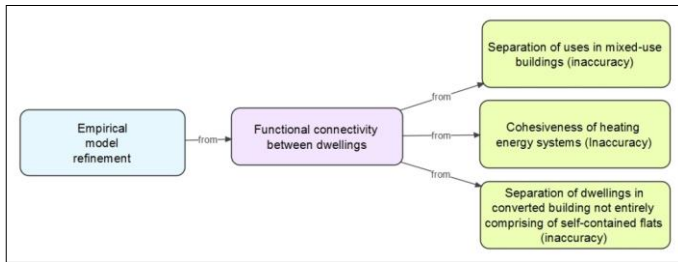


Fig. 9 Uncertainties in the empirical model refinement

Issues arise at: (i) separation of uses in mixed-use building (inaccuracy), as the criteria is to separate uses vertically in different buildings. This approach seems to work well with mixed residential/commercial tenures in Newcastle residential buildings; (ii) the cohesiveness (inaccuracy) in identifying the district and group heating infrastructures, and (iii) the separation of dwellings (inaccuracy) in converted buildings not entirely comprised of self-contained flats. The inaccuracy results from some buildings being misclassified in their type, their fuel type and their use. This misclassification can result in a group of dwellings being incorrectly classed residential (or commercial) and their annual estimated energy consumption being incorrectly estimated with values for shared amenities (or not). This contributes to uncertainty in the NCRF.

*G. DECC data set level*

Issues from the best DECC data set level follow the concept map segment in Fig. 10.
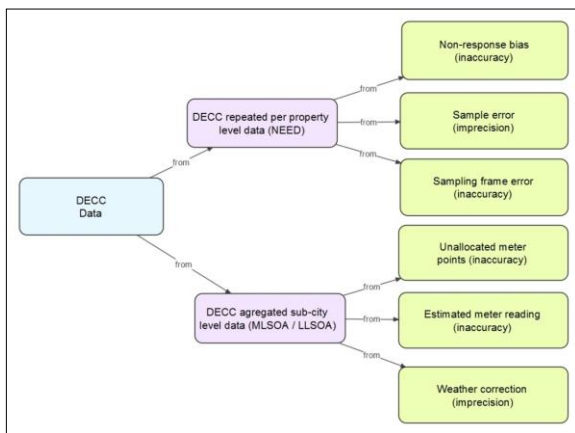


Fig. 10 Uncertainties in the DEC data set level

i. Repeated Heating by property attributes NEED

In NEED, a systematic error (inaccuracy), or bias can occur, for example, when: (i) non-metered fuels e.g. oil or coal is omitted from the sampling frame, and (ii) excluded properties may have only used gas for other purposes, such as cooking.

Other sources of inaccuracy are: (i) only households with valid heating gas consumption, between 2,500kWh and 50,000kWh, have been included; (ii) all consumption figures are rounded to the nearest 100kWh, and (iii) any estimates based on fewer than 30 properties have been excluded from the tables.

The most important component of random error in the NEED data set is sampling error (imprecision), which is the error that arises because the estimate is based on a sample survey rather than a full census of the population. Four million records (i.e. a 17.8 per cent sample of UK housing stock) were drawn from the VOA database (DECC, 2014, p. 28).

ii. Aggregated MLSOA and LLSOA domestic electricity and gas estimates.

In DECC, systematic error (inaccuracy) or bias occurs in the following cases: (i) DECC annualized consumption data for the Meter Point Administration Number (MPAN) or electricity meter data used in the analysis consists of approximately 80 per cent actual ("Annual Advance") readings and 20 per cent estimated readings ("Estimated Annual Consumption"); (ii) DECC electricity consumption data for each MPAN is not weather corrected, and (iii) the sum of meter points or domestic energy consumption at MLSOA level does not always equal the sum of meter points or domestic energy consumption at the associated LA level. Similarly, the sum of meter points or domestic energy consumption at the LLSOA level does not always equal the sum of meter points of domestic energy consumption at the associated MLSOA level due to unallocated meters.

In summary, this section has proposed a concept map to quantify the uncertainties in the model outcome in line with the issues associated with three sources –the model, the full input SAP and the refinement/validation issues.

## III. SUMMARY

The uncertainty taxonomy from Sections I and II can be summarized according to few dominant factors:

i. The underlying assumptions about processes exogenous to the model (e.g. climate variables).This will reflect in regional average parameters to be entered into the model.

ii. The underlying assumptions about endogenous processes in the model (e.g. spatial interpolation). Because none of the three case study zones are homogeneous, the resulting algorithm produces different surface structures in Westgate, which is a complex area.

iii. The assumption in judgements, such as the ontology of the CHM model, e.g. in the building occupant behaviour, cannot only be a function of the usable floor area.

iv. The simplifying assumptions in the structure, e.g. the sum of meter points or domestic energy consumption at the LLSOA level does not always equal the sum of meter

points of domestic energy consumption at the associated MLSOA level due to unallocated meters.

v. Finally, as explained, different weather correction methodologies in DECC/NEED and CHM lead to some discrepancies (inaccuracy), which probably vary somewhat from year to year, because the exact methodology for weather correction for NEED/DECC is not fully disclosed.

The concept map lays out all these factors in a common diagram, so that NCRF can be better understood and synchronized in other cities.

## ACKNOWLEDGMENT

## REFERENCES

BRE (2011) *The Government's Standard Assessment Procedure for Energy Rating of Dwellings. 2009 edition incorporating RdSAP 2009* (rev October 2010, RdSAP 2009 added March 2011). Garston, Watford: BRE. [Online]. Available at: http://www.bre.co.uk/filelibrary/SAP/2009/SAP-2009_9-90.pdf (Accessed: 05 November 2013).

Calderón, C., James, P., Alderson, D., McLoughlin, A. and Wagner, T. (2012) 'Data availability and repeatability for urban carbon modelling: a CarbonRouteMap for Newcastle upon Tyne', *Retrofit 2012*. Manchester England. Available at: http://www.energy.salford.ac.uk/page/day_2_papers (Accessed: 3 March 2012).

Calderón, C., James, P., Urquizo, J. and McLoughlin, A. (2015) 'A GIS domestic building framework to estimate energy end-use demand in UK sub-city areas', *Energy and Buildings*, 96(0), pp. 236-250.

DECC (2013) *The Fuel Poverty Statistics Methodology and User Manual* (URN 13D/215). London: DECC. [Online]. Available at: https://www.gov.uk/government/publications/fuel-poverty-methodology-handbook-2013 (Accessed: 09 December 2013).

DECC (2014) *National Energy Efficiency Data-Framework. Summary of analysis using the National Energy Efficiency Data-Framework (NEED)* (URN 14D/192). London: DECC. [Online]. Available at: https://www.gov.uk/government/collections/national-energy-efficiency-data-need-framework.

Han, P.K.J., Klein, W.M.P. and Arora, N.K. (2011) 'Varieties of Uncertainty in Health Care: A Conceptual Taxonomy', *Medical Decision Making*, 31(6), pp. 828-838.

Hughes, M., Palmer, J., Cheng, V. and Shipworth, D. (2013) 'Sensitivity and uncertainty analysis of England's housing energy model', *Building Research & Information*, 41(2), pp. 156-167.

Novak, J.D. and Cañas, A.J. (2006) *The theory underlying Concept Maps and how to construct them*. Florida Institute for Human and Machine Cognition. [Online]. Available at: http://cmap.ihmc.us/Publications/ResearchPapers/.

Shapiro, S.C. and Eckroth, D. (1987) *Encyclopedia of artificial intelligence*. Wiley.

Sowa, J.F. (2006) 'Semantic Networks', in *Encyclopedia of Cognitive Science*. John Wiley & Sons, Ltd.

Urquizo, J., Calderón, C. and James, P. (2016) *2016 IEEE Ecuador Technical Chapters Meeting (ETCM)*. 12-14 Oct. 2016.

Urquizo, J., Calderón, C. and James, P. (2017) 'Metrics of urban morphology and their impact on energy consumption: A case study in the United Kingdom', *Energy Research & Social Science*, 32(Supplement C), pp. 193-206.

Urquizo, J., Calderón, C. and James, P. (2018a) 'Modelling household spatial energy intensity consumption patterns for building envelopes, heating systems and temperature controls in cities', *Applied Energy*, 226, pp. 670-681.

Urquizo, J., Calderón, C. and James, P. (2018b) 'Understanding the complexities of domestic energy reductions in cities: Integrating data sets generally available in the United Kingdom's local authorities', *Cities*.