Urquizo J, Calderón C, James P.

Understanding the complexities of domestic energy reductions in cities:
Integrating data sets generally available in the United Kingdom's local
authorities.

Cities 2018

DOI: <https://doi.org/10.1016/j.cities.2017.12.019>

Copyright:

© 2018. This manuscript version is made available under the [CC-BY-NC-ND 4.0 license](#)

DOI link to article:

<https://doi.org/10.1016/j.cities.2017.12.019>

Date deposited:

05/01/2018

Embargo release date:

30 June 2019

This work is licensed under a

[Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International licence](#)

Understanding the complexities of domestic energy reductions in cities: Integrating data sets generally available in the United Kingdom's local authorities

Javier Urquizo^{a,c,1}, Carlos Calderón^a, Philip James^b

^aSchool of Architecture Planning & Landscape, Newcastle University
Newcastle upon Tyne, UK, NE1 7RU

^bSchool of Civil Engineering & Geosciences, Newcastle University
Newcastle upon Tyne, UK, NE1 7RU

^cEscuela Superior Politécnica del Litoral, Facultad de Ingeniería en Electricidad y Computación,
Campus Gustavo Galindo Km. 30.5 Vía Perimetral,
P.O. Box 09-01-5863, Guayaquil, Ecuador

Abstract

This paper describes the development of a spatial database for an annual energy consumption framework. As the database of the existing building stock is a weak point in most European cities but possible pathways to energy reduction in the building sector have to be found, this paper is of high relevance. Our framework uses a complex multi-source data set and a plethora of statistical methods to merge various large complex data sets and then apply a heat balance model in three sub-city areas in Newcastle upon Tyne, United Kingdom. The framework estimates the energy end-use at the single dwelling level on three aggregated scales: district, neighbourhood and communities. We propose a methodology for modelling energy in buildings in different ways depending on the required output scale, the cluster top-down model and the domestic energy model (DEM) bottom-up approaches. The cluster model is a generalization of similar building energy profiles into archetype medoid prototypes in districts and eventually the whole city, whereas the sub-city DEM is a representation of an individual building energy profile in neighbourhoods and communities. The framework can be used to test different energy measures (fabric and heating supply systems) for the same property type, and give insights into community energy analysis by aggregating individual building energy consumption. These insights will enable rational and considered responses to be formulated to the problems of integrating renewables into the generation portfolio, that are likely to be faced

¹ Corresponding author

Email addresses: jurquizo@espol.edu.ec (Javier Urquizo), carlos.calderon@newcastle.ac.uk (Carlos Calderón), philip.james@newcastle.ac.uk (Philip James)

in the future.

Keywords: cross-study analysis, data reduction, domestic energy consumption, energy mapping, model refinement, record augmentation, spatial interpolation, structure detection, spatial urban energy modelling.

1. Introduction

After a series of European policy initiatives (e.g. the European Strategic Energy Technology Plan and the Partnership for Energy Efficiency Cooperation), European cities are now in need to ensure that new buildings, as well as large existing buildings meet certain minimum energy requirements. It also requires that all buildings should undergo 'energy certification' prior to sale, and that boilers and air conditioning equipment should be regularly inspected. However, a unique future national energy scenario would translate into many alternative energy futures pathways at the local authority (LA) scale. As a result, LAs require a spatial understanding of end-use energy demands in targeted city areas to provide evidence for meeting future challenges in planning local energy services and infrastructure. The geographical coverage is closely associated with the purpose, i.e. a descriptive energy model could draw a baseline for the city energy consumption. In this energy modelling research approach, the concept of spatial data integration ensures multi-sourced data and interoperability. The information can be integrated through key identifiers and the result is a comprehensive, spatially enabled database in the energy model. A robust spatial and modelling methodology would enhance the effectiveness of their local energy planning of future interventions of energy efficiency measures, decentralised energy supply infrastructure. For instance, in each city of the United Kingdom (UK) some local areas might be more suited for building envelope technological interventions such as those promoted by the Green Deal whereas others may benefit from centralised and distributed energy supply options.

The literature on energy system frameworks involves different (sometimes

linked) approaches for the purpose of decision analysis. Examples are: Ascione et al. (2013) use a combined method of already available information and energy audits on a sample of buildings to quantitatively estimate the energy consumption for each building. Trigaux et al. (2017) use a design tool that extracts solar obstruction data, building geometry and constructing elements, and then uses this information to assess the heating energy demand based on the dynamic Equivalent Heating Degree Day (dEHDD) method. Ma and Cheng (2016) use a Big Data methodology (Hsu, 2015) for integrating a large number of predictors and a Geospatial Regression method for filling the blanks in the data set. The Ma and Cheng method then rank the features and estimate the energy use in a single building using a case study in New York City. Vartholomaios (2017) uses a standardised rank regression coefficient sensitivity technique on a parametric study of three urban topologies to understand the co-influence of the urban form parameters on the energy consumption. Yeo et al. (2013) use an Energy Integrated Planning Support System to model urban space, which predicts urban micro-climate and energy demands. The Yeo et al. model includes a function to ultimately visualize the 2D and 3D information for energy planning in the city. In the United Kingdom, Rylatt et al. (2003) propose an innovative use of Geographical Information System for Estimating Municipal Energy for Residences using Arbitrary Levels of Data (EMERALD) based in the Building Research Establishment Domestic Energy Model (BREDEM) (Anderson and Chapman, 2010) that produce monthly predictions of annual energy use for space and water heating, lights, cooling and appliances.

Our paper argues that energy planning at decentralized level would be to prepare an area-based DTS to meet energy needs using a bottom-up approach and disaggregated data (Hiremath et al., 2007, p. 735). This paper presents a framework to estimate spatial domestic energy consumption in individual dwellings and aggregated areas using a spatially enabled database to represent local area characteristics. We use a cluster framework for districts and a sub-city DEM framework for neighbourhoods and communities, the Newcastle CarbonRoute Framework (NCRF) database. NCRF is a spatially referenced parameterised per-dwelling domestic energy framework developed with the purpose of estimating the energy consumption of sub-city areas. Two types of modelling were carried

out. In the first, top-down, using K-means clustering of dwelling archetypes and the second a sub-city Domestic Energy Model (DEM) that creates individual energy consumption estimates for each dwelling and aggregates these to sub-city areas.

Principal component analysis underpinning the cluster model was used to provide both a method for explaining how many of the surveyed categorical variables are needed to explain the total variance within the reduced data set, and to detect the correlations between the variables. Besides these more obvious uses, this study has shown that there is an additional benefit as the principal component also provides group estimates of English Housing Survey 'donor' records before the Newcastle CarbonRoute Map (Calderón et al., 2012) nearest neighbour multiple imputations method actually selects the best 'donor' in an augmentation strategy for inputting to a BREDEM type model. This shows that the principal component analysis is a worthwhile process; however, it remains to be seen whether the principal component results will be transferable to other cities because regional climates make for different insulation and heating supply requirements. However, as this paper has demonstrated the NCRM data sets are generally available in some form to many, if not all, local authorities and other local authorities should consider this type of approach to collate data from existing sources with the requirement that there has to be a large sample so its application will be statistically valid.

The NCRF is likely to be of significant value as a policy tool. It can be used to inform and direct policy by testing the effect that various policy decisions are likely to have on the community energy consumption. For instance, the framework could be used to test the effect of possible future revisions to the Building Regulations or energy efficiency schemes. We believed that current UK Government regional and sub-city methods and data for domestic properties in its current most disaggregated form may not accurately represent energy consumption of geographically specific and homogenous urban areas in the UK and therefore be insufficient for providing evidence for meeting future challenges in planning local energy services and infrastructure; also, NCRF will significantly increase our theoretical understanding of the complex interrelationships that exist, not only

between the various end-uses of energy within the demand side of the UK housing stock, but also the relationships that exist between the demand side and a particular combination of urban form, climate, buildings physical characteristics and behaviour and household social practice (Shove and Walker, 2014) in a sub-city area.

From Figure 1, this paper is organized as follows: Section 2 is a brief review of the initial data sets. Section 3 explains the procedures to integrate the local and the national data sets to produce an augmented Newcastle CarbonRoute Framework (NCRF). Section 4 explains the NCRF model refinement. Section 5 uses the Cambridge House Model (CHM) to estimate annual electricity and gas consumption and finally Section 6 makes the discussion and the conclusions and the concluding remarks are in Section 7.

The software components used in NCRF are in two different contexts and two kinds: (i) using components that are parts of a commercial executable e.g. the spatial interpolation components from the ArcMapTM from ESRI or CHM, or (ii) using executable modules custom made for this study, and called Structured Query Language (SQL) scripts. As a summary, Figure 2 is the main computing process flow leading to the NCRF energy consumption estimates.

The diverse nature of the inputs into the NCRF has resulted in a significant library of PL/SQL scripts used for cleaning, sorting, importing and linking NCRM data. The modular nature of the development means it is relatively easy to add new data sets or update existing ones using the scripts developed. In addition, as the energy model is integrated only loosely with the NCRF, other energy models could be utilised if appropriate either in a similar fashion to CHM though matching energy results to records or through directly linking NCRM records with an executable or script.

The final output of the NCRF is a spatially enabled dataset of residential dwellings with a per-dwelling energy estimate and a large number of energy-related variables. PostGIS^{T M} tables can be linked to many spatial software or data can be exported directly so that further analysis or visualisation processes can be carried out. In this study ArcMapTM was used to carry out

further analysis of the NCRF results.

Figure 1: NCRF energy estimates roadmap

Figure 2: NCRF process flow

2. A brief review of the initial data sets

The aim of this paper is to create an energy framework at building level that can be utilized with minimum data collection, just enough to calculate the energy consumption from different house types representing the whole city, in our case study the Newcastle domestic stock; for this purpose, a large number of local building and dwelling data were combined, the result is the

Newcastle CarbonRouteMap (NCRM) (Calderón et al., 2012) and ultimately linked to the national English House Survey (EHS).

This paper will use the term 'NCRM cluster profile' when referring to a dwelling having these six combined variables in its profile (Palmer and Cooper, 2012, p. 21), this is the dwelling archetype for the cluster model whereas the term 'NCRM DEM profile' will be used when referring to a dwelling having ten variables in its profile. The term 'NCRM Standard Assessment Procedure (SAP) profile' is reserved when a dwelling has 115 variables (a full SAP) in its profile. Based on the selection of different profile characteristics in Newcastle buildings, different energy consumption is expected.

This study has developed a spatial database for the purpose of understanding the energy consumption of Newcastle upon Tyne at building level; this data set provides the six category variables (the basic building profile). The descriptive statistics of the NCRM sample is in Appendix A, and the formation of NCRM is described in Appendix B (the NCRM level envelope parameters and tenure) and C (the NCRM level fuel supply for heating systems). For each building form in NCRM, additional typical parameters from the dwellings they represent in the EHS data set were included. Thus, parameters such as wall areas, glazing, hot water heating, secondary heating fuel, etc. were derived from real buildings in the EHS to form the NCRM SAP profile. The EHS is the second major data set in this study and is described in Appendix D.

In general, houses (66%) are the most common built form in Newcastle; whereas flats (24%) are increasing due to the large and growing student population needing accommodation near universities. Bungalows and maisonettes are less common types. In the house segment, the semi-detached and terraced houses are the most common detachment in

Newcastle (each representing 43% and 45% of house stock respectively), whereas in the flats segment, it is common to find Tyneside flats, which are a pair of single-storey purpose-built flats within a two storey Victorian terrace.

3. NCRM Methodology

Newcastle Carbon Route Map Framework (NCRF) is a spatially referenced parameterised per-dwelling domestic energy framework developed with the purpose of estimating the energy consumption of sub-city areas. The second modelling process is a sub-city Domestic Energy Model (DEM) that creates individual energy consumption estimates for each dwelling and aggregates these to sub-city areas. Both models utilise a physic based approach to energy modelling based on BREDEM 12 methodology. In our case we use the Cambridge Housing Model (CHM) which implements a BREDEM based energy model to compute energy estimates as it is the model used by DECC to underpin the 2012 Housing Energy Fact File and Energy Consumption in the United Kingdom. The NCRF sub-city DEM modelling process is broken down into several separate stages.

3.1. Cross-study analysis to integrate NCRM and EHS data sets

Cross-study analysis is a technique for creating a statistical combination of two separate studies. The whole strategy is to have for each individual NCRM record an augmented record to run a Building Research Establishment Domestic Energy Model BREDEM-like energy model, therefore, to be able to compute the energy consumption for an individual dwelling. The record augmentation technique uses near neighbour multiple imputation to find an appropriate EHS record to extend each NCRM record. Cross-study analysis studies poses two statistical challenges that need to be overcome: What were the survey sample designs? and second, do the survey questions and response categories match?

Section 3.1 describes the survey sample design and Section 3.2 describes the matching of the categories and the definitions of the responses, known as harmonization. This paper performs two types of harmonization: semantic and cross-scale.

3.2. Sample survey design for cross-study normalization

There are two potential approaches to cross-study normalization depending on whether the information being synthesized is either higher level or sample level. At the highest level, the cross-study normalization approach is to combine the studies (Espey and Espey, 2004), where the broad conclusions of different studies are synthesized. However, (Lin, 2012) argues that the danger of this approach is in structuring a large set of different studies, as the assembled definitions can become imprecise and the results may be difficult to interpret meaningfully.

We argue that a more suitable approach is to consider directly standardized sample-level measurements within each study. The published literature (DCLG, 2010, pp. 85 - 88) on the EHS sampling and grossing methodology states that the full EHS household data set has a stratified random sample design and the dwelling or paired (one per each year of the survey) data set has been stratified according to tenure type (Rafferty, 2011) (rented tenures are over-represented and owner-occupation under-represented in relation to national estimates). The EHS survey of the housing stock (DCLG, 2011, p. 71) is effectively a simple random sample.

To prove that the NCRM is also a stratified random sample, we use cluster analysis. A cluster analysis result needs to show that the resulting NCRM samples by Middle Layer Super Output Areas (MLSOAs) are distributed in the same way as the whole population (Newcastle city) in terms of the stratifying criteria (of the six energy predictors: dwelling size, dwelling age, building form, wall construction, heating and number of

floors). If both studies (NCRM and EHS) are stratified random samples, then it is possible to merge these into a single data set.

For the NCRM data set, the sample-level measurement applied is the spatial cluster technique with the following rationale. In order to have a gross-error-free sample ready for cluster analysis, a process was applied to remove records that contain null, 'do not know', or zero values in any field. The final count of the NCRM residential polygons in the sample is made up of 60,977 (out of 78,475 dwelling polygon features). This new Newcastle sample has the dwelling-records with a sampling centre in the sample median. Treating the case as a cluster and quantitatively measuring its relevant dimensions allow us to systematically compare it with other cases, and therefore to address the external validity problems that are common to the single case study design (Stake, 1978). Also, another benefit from formally sampling units of analysis within a case cluster is the possibility for cross-stratification with different data sources in order to sort out the effects of measurement or perspective (McClintock et al., 1979).

A hybrid method was applied to determine the number of case clusters and its composition. First, hierarchical clustering (the Wards method) identified the eight homogenous groups of house archetypes (clusters), where group assignments were updated iteratively until convergence with a local minimum of the sum of squared Euclidean distances was reached (Carlsson and Mémoli, 2010), as seen from the dendrogram. The procedure involves considering whether various sub clusters make sense from the point of the intuitive (dis)similarities between data and the scope of the investigation (Robinson, 2014). Then, K-means cluster analysis (Burns and Burns, 2008) was applied over this NCRM gross-error-free sample with only the chosen optimum number (8) of clusters in which to place all the observations.

These eight meaningful case clusters are cluster 1 (18.24%); cluster 2 (0.2%); cluster 3 (1.71%); cluster 4 (9.98%); cluster 5 (30.98%); cluster 6 (12.18%); cluster 7 (22.67%), and cluster 8 (4.04%). As an example, uninsulated houses are in cluster 4, and all-insulated houses are in cluster 6; low efficient heating systems are in cluster 2 and high efficient heating systems are in cluster 3. Furthermore, this case cluster classification is applied to MLSOA level to reveal information about the underlying energy consumption in districts. This potentially generates spatial pattern distributions of domestic energy use, which can be of further aid in understanding the house archetypes and their relationship with energy consumption.

In summary, both the EHS sample and the NCRM gross-error-free sample are stratified random samples, and can then be integrated to form individual dwelling full SAP registers, provided the harmonization procedure assures similar output responses (semantically and cross-scale) as shown in Section 3.3.

3.3. Output response harmonization

In recent years, there has been an increasing amount of literature on harmonization and in general two approaches can be identified. Input harmonization aims to achieve standardized measurement, processes and methods in all national or regional populations, while output harmonization uses different national and local measurements, mapped into a unified measurement scheme (Granda and Blasczyk, 2011). Output harmonization techniques (semantic and cross-scale) were used to map NCRM to EHS categories in each of their responses.

i. Semantic output harmonization.

Semantic output response harmonization procedures were applied on

NCRM to ensure that all the variables are named the same across both data sets (NCRM and EHS) and that they refer to the same attribute; also, the NCRM information scattered over several variables was consolidated into a single record (NCRM Cities Revealed, NCRM SCORCHIO, NCRM Your Homes Newcastle and NCRM Registered Social Landlords).

ii. Cross-scale output harmonization

Cross-scale harmonization is used to combine data sets from different scales. In this study, cross-scale harmonization is applied to the floor area definitions in the local and national data sets. There are several ways of defining floor area (Washington State University, 2007): (i) the total floor area contained within the building measured to the external face of the external walls or gross floor area (GFA); (ii) the floor area contained within the building measured to the internal face of the external walls or gross internal area (GIA); and (iii) the net internal area (or usable floor area). While NCRM uses the whole building footprint GFA, EHS (DCLG, 2009) uses the usable floor area of individual self-contained dwellings.

This paper uses a three-stage method to find the individual dwelling usable floor area (dwelling size) in NCRM data set. Each individual dwelling usable floor area was corrected to its actual value from the original Topographic Identifier footprint area by a sequential process involving the use of secondary information found in three sources. First, the building type classification from the SCORCHIO project (Triantakou and Barr, 2009) identifies the number of residential and commercial properties per building, and then the addressing conventions method in LGIH (2010) was used to indicate the vertical position of dwellings in relation to ground level in complex and grouped property types from the Gazetteer. This was done to find the number of

residential dwellings per storey and the number of storeys in the building. These two stages allow the individual usable dwelling floor area to be found by fitting the dwellings in the same storey to the building foot print area. In the third stage, the housing market information validates the two previous steps with a final approximate individual usable floor area for each dwelling.

To take into account intra- and inter-regional differences, cross-scale harmonization was prepared according to the procedure used by Lorimer (2010). In each MLSOA, the national computed floor area is computed as shown in Equation 1.

$$a_f = a_i * \left(\frac{m_{na}}{m_{la}} \right) \quad (1)$$

Where a_f is the 'national computed area'; a_i is the 'local computed area'; m_{la} is the mean GFA area based on residential density by type in Newcastle local authority and m_{na} is the mean GIA area for each dwelling type by region and area type (as reported by the EHS surveyor).

Issues from cross-scale normalization and harmonization considered in this paper were: (i) semantic interoperability. i.e. 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. (ii) floor area definition must be allocated

for every dwelling from information available for number of storeys, height and footprint; this is estimated in mixed-use and multiple storey buildings.

3.4. NCRM data reduction and structure detection

The cross-study analysis was a necessary pre-processing step for NCRM and EHS integration; however, for the actual NCRM record augmentation a procedure was required for the identification of the building parameters with the strongest physical determinant used to model the house reduced archetype. This was accomplished by the Principal Components Analysis (PCA) (reducing the data to a smaller number of components) (Everitt et al., 2011, p. 29), and Factor Analysis (FA). FA was used to understand which constructs underlie the NCRM archetype parameters (Everitt et al., 2011, p. 57). The PCA and FA help to reduce the number of records of the target EHS data set to only those possible records that could potentially augment NCRM records to a full SAP record (the extended archetype). The procedure to perform the PCA has two steps; first is the computation of the correlation matrix, and then the share of variance of the original data set is explained by each principal component. The approach is to determine the extent to which each variable is explained by each dimension (or PC). The final result of the FA is a matrix that indicates which variables (and how much) saturate (correlate) each of the factors found. Saturation represents the weight or importance of the variable within the PC (or dimension).

Results from the correlation matrix show higher correlation (association) between dwelling size, building form and number of floors (principal component one PC_1); and between age and wall construction (principal component two PC_2). The total variance table informs the cumulative proportion of variance criteria that can be met with two components. Also, PC_1 and PC_2 satisfy the criteria of explaining 60% or more of the total variance.

3.5. *Case study selection*

Three case studies were selected based on different urban morphology. The urban morphology refers to the spatial configuration of urban land use within an urban area. Urban morphology has profound influences on the energy consumption of a city (Urquiza et al., 2017) and can affect the residential energy use through three causal pathways: directly through electric transmission and distribution losses, and indirectly through the housing stock and the formation of urban heat islands (Ewing and Rong, 2008). This paper is concerned with the indirect impacts. Three metrics are used to estimate the implications on the energy consumption.

i. Absolute Settlement patterns

This paper has selected three MLSOA that have interesting settlement patterns: South Heaton, Westgate, and Castle (Urquiza et al., 2016), see Figure 3. Coombes and Raybould (2001, p. 224) argue that there are at least three key dimensions to consider in modern settlement patterns: size, concentration, and accessibility.

Figure 3: Newcastle MLSOAs: Castle, South Heaton and Westgate

South Heaton has an area of 102 hectares (the smallest in the city) and shows a high concentration of late Victorian/Edwardian terraced houses, and has become a very popular residence for students, and this is reflected in the high number of Houses in Multiple Occupation (HMO) (nearly 300). Westgate has an area of 501 hectares and contains the commercial centre of the city and has a high concentration of residential self-included dwellings and mix-use buildings. Castle has an area of 2,179 hectares (the biggest in the city); the sixties/seventies semi-detached and detached houses are the most important house types.

ii. Spatial Urban Intensity patterns

Stone and Rodgers (2001) argue that spatial Urban Intensity is used to describe the overall condition of residential energy land-use intensities

i.e. high density urban areas (usually city centre) alter their climate in the form of elevated temperatures relative to suburban (usually the peripheral areas). Castle is the lowest density (1.8 dwellings per hectare) district in the city and South Heaton is the largest (43.48 dwellings per hectare).

iii. Socio-economic Urban patterns

Since 2009, the City Council has been able to identify and attain evidence of where some of the most vulnerable groups within the city are located. The data is provided by (Rowntree, 2010) in the form of a number and a percentage of households within each of the broad socio-economic classification groups and types. Experian data shows that Castle has middle income younger families living in modern houses; South Heaton has students, young couples and singles living in small, old flats, and Westgate has low income persons living in social housing and young, educated persons living in temporary accommodation.

3.6. *Spatial interpolation methods for asserting parameters*

This paper uses the multiple case study approach to test the spatial interpolation strategies to create complete SAP spatial records for energy analysis below the local authority level. Spatial interpolation methods are one of the more practical means of group-based estimation. The rationale behind spatial interpolation is the observation that points close together in space are more likely to have similar values than points far apart, Tobler's Law of Geography, (Sui, 2004). A basic assumption is that the value to be estimated at a certain point is influenced more by nearby points than those that are further away (Chang, 2012). For the known points to be effective, they should be well distributed within the case study area.

A variety of spatial interpolation methods (Li and Heap, 2008) exists.

This study uses three different interpolation methods to estimate the dwelling parameters: nearest neighbour (NN), inverse distance weighting (IDW) and kriging. All methods are weighted average methods, and they all have the same basic mathematical formulation to compute a dwelling parameter z at an unsampled point x_0 , given a set of neighbouring sampled values z_k , and sampled at locations denoted by x_k . The interpolating relationship is shown in Equation 2.

$$z(x_0) = \sum_{k=0}^n \lambda_k z(x_k) \text{ and } \sum_{k=0}^n \lambda_k = 1 \quad (2)$$

Where λ_k represents the weights assigned to each of the neighbouring values, and the sum of the weights is one. For these approaches, interpolation involves three steps: defining the search area or neighbourhood around the point to be predicted; locating the observed data points within this neighbourhood (Stone and Rodgers, 2001) and assigning appropriate weights to each of the observed data points. Neighbouring strategies were used in Castle and South Heaton.

In complex neighbourhoods like Westgate in which there are tall buildings, the kriging interpolation method was used. The variogram (Cressie, 1990) is essential for kriging and plays a useful role as the function that describes spatial dependence for an area (geo-referenced) variable. It assumes that the spatial variation in the phenomenon is statistically homogeneous throughout the surface; that is, the same pattern of variation can be observed at all locations on the surface. Also kriging automatically accounts for information redundancy, thus eliminating the need for de-clustering corrections (Caers, 2011, p. 103).

The urban form characteristics suggest the best method to adopt. For a high compact area like South Heaton, this is the Nearest Neighbour (NN) method. NN selects only a single sampled value closest to the point of

interest and assigns it a weight of one; NN was selected on the basis of a high compact residential dwelling area and sample penetration. For Castle, where the residential property is scattered, the preferred interpolation is Inverse Distance Weighting (IDW). As observed values that were closer to the point of interest were more heavily weighted, IDW was selected on the basis of low residential density. For complex Westgate ordinary kriging was used as it helps to compensate for the effects of data clustering, assigning individual points within a cluster (residential self-included dwellings in the same building) less weight than isolated data points, i.e. treating clusters as single points. The spatial interpolation process generated full spatial data sets on Castle, South Heaton and Westgate (that is, all residential properties are assigned the six key energy variables).

3.7. NCRM record augmentation strategy

The record augmentation is creating a full SAP record through nearest neighbour imputation that completes the NCRM record with information from EHS records. The BREDEM-8 model is used in CHM. CHM requires an input record containing 115 variables based on a full SAP survey. In order to generate this for a NCRM record, the key attributes identified through the PCA and FA are used to find a matching record in the EHS and append these values to the NCRM record. In addition to the six physical predictors, tenure is also used. It has been argued (Pattison et al., 2010, pp. 8 - 10) that important differences in stock condition and energy efficiency can be identified between the three main tenures in the UK i.e. owner occupation, private renting and social renting. The full set of common fusion variables (CFV) (Dorofeev and Grant, 2006, p. 216) for the imputation procedure is shown in Table 1. The term fusion is applied to the linking or merging of data sets that have common elements (variables). Data fusion is the process of fusing multiple records representing the same real-

world object into a single, consistent, and clean representation (Bleiholder and Naumann, 2009). This process is also referred to in the literature as data merging or data consolidation.

CFV code	Description
floorarea	Usable floor area
dwtype7x	Dwelling type
fodconst	Construction date
storeyx	Number of floors above ground
typewstr	Predominant type of wall structure
Felcavff	Cavity wall insulation
finmhue	Main heating fuel
Finchtyp	Primary heating system - type of system
finmhboi	Boiler group
tenure8x	Tenure

Table 1. Common fusion variables

The procedure works in two stages, the first of which is creating the NCRM physical field (age, infrastructure and land use) for grouping similar individual dwellings, and the second is applying the nearest neighbour imputation procedure to find for each dwelling in the NCRM data set the best record within the EHS data set.

The nearest neighbour imputation algorithm works as follows: if the NCRM data set ($a_1, a_2 \dots a_{60,977}$) has 60,977 records, each with $p = 10$ variables (the common fusion variables) that is 60,977 records in a p -dimensional space, and the EHS data set ($b_1, b_2 \dots b_{16,150}$) has $q = p + 95 = 105$ variables or 16,150 records in a q -dimensional space; then the nearest neighbour imputation will find a set of plausible EHS records that extend NCRM from a p -dimensional space to a q -dimensional space, where the geometric property (Jolliffe, 2002) of the selected EHS record(s) is minimizing the sum of the weighted distances $w_i |a_i - b_i|$ for all i in the CFV (see Equation 3). The appropriate subspaces are traversed subsequently, first in the PC_1 -dimensional subspace, then the $(PC_1 + PC_2)$ -dimensional subspace and then the full ten-dimensional subspace of the CFV. The

possible outcomes of these subspaces are explained in the next paragraph.

The imputation procedure is summarized in Figure 4. First, for each dwelling in NCRM it finds a set of plausible EHS Records ($SPER_1$) that have the same values in the variables of the PC_1 (dwelling size, dwelling form and number of floors); a second iteration reduces the number of EHS records in $SPER_1$ to those EHS records that additionally, for each dwelling in NCRM, have the same values in the variables of the PC_2 (age, predominant type of wall structure and cavity wall insulation), a new set of plausible records is found ($SPER_2$). A third iteration reduces the number of plausible EHS records in $SPER_2$ even more using the same values for heating (main heating fuel, primary heating system and boiler group), a new set of EHS records is found ($SPER_3$); and finally the procedure is repeated with the value of tenure. The final outcome is a number of plausible EHS records ($SPER_4$) for each dwelling in NCRM can be null (zero) ('not exact'), one ('key term') or more than one record ('exception').

Figure 4: NCRM record augmentation strategy

In summary, for each record of interest in NCRM, the algorithm will identify candidate records in EHS that match the record of interest. It has three outcomes: (i) ‘exceptions’, this is used when a unique or best suitable EHS candidate record cannot be found, since several records look equally valid where several EHS records have the same values in all CFV; (ii) ‘exact match’, this is the perfect outcome, where there is a single best suitable EHS record with all CFV being equal in both data sets; and, (iii) ‘not exact match’, this uses minimal Weighted Distance (minWHD) of dimension 10, to find the most appropriate EHS record, as shown in Equation 3.

$$\text{minWHD} = \sum_{i=0}^{10} w_i |a_i - b_i| \quad (3)$$

Where a_i and b_i are the i_{th} -dimension subspace arguments of the data sets NCRM and EHS respectively and $w_i \in$ between 0 and 1 are the associated weights.

3.8. Energy consumption calculation

The core per-dwelling energy calculation approach in the UK is a building physics one. In the UK, the most widely used building physics-based models for energy consumption share the same BREDEM core calculation engine (Kavgic et al., 2010, p. 1685). BREDEM uses a mixture of analytical and empirical functions. It was decided that the best BREDEM-like model to adopt for this paper is the Cambridge Housing Model (CHM); the calculations in the CHM are principally based on the worksheet in SAP 2009, the Government’s Standard Assessment Procedure for energy rating of dwellings, plus the Reduced Data SAP for existing dwellings (BRE, 2009, pp. 99 - 121). The SAP 2009 outputs for energy use and associated CO₂ emissions, and do not include cooking or electrical appliances. CHM has therefore included calculations for energy use for cooking and electrical appliances, and

associated CO₂ emissions, based on the BREDEM-8 and SAP, see Figure 5. However, our framework approach could be applied to any other energy model.

Figure 5: CHM model design (Hughes, 2011, pp. 1)

CHM (Hughes et al., 2012) uses the extended (complete) SAP NCRM dwelling profile with information on: (i) dwelling and household information; (ii) geometry; (iii) ventilation; (iv) other heat loss elements; (v) space heating; (vi) hot water system; (vii) low energy lighting as the input record. CHM runs were performed on Castle, South Heaton and Westgate. Each complete NCRM dwelling consists of 115 energy variables, as shown in Table 2 and Equation 4.

Dwelling and Household information	Geometry	Ventilation	Other heat loss elements	Space Heating	Hot water system	Low energy lighting
H ₁ ...H ₉	H ₁₀ ...H ₄₂	H ₄₃ ... H ₅₆	H ₅₆ ... H ₈₄	H ₈₅ ...H ₉₇	H ₉₈ ...H ₁₁₄	H ₁₁₅
Dwelling: code, region, tenure, type, age. Household: Number of occupants (adults and children)	thermal insulation of the building fabric; materials used for construction of the dwelling	characteristics of the ventilation equipment	ventilation characteristics of the dwelling	the fuel used to provide primary and secondary space heating; efficiency and control of the heating system	the fuel used to provide water heating; renewable energy technologies	the fuel used to provide lighting

Table 2. NCRM SAP input

$$H_j = (H_1, H_2 \dots H_{115}) \quad (4)$$

The CHM reads data for each NCRM dwelling, performs building physics calculations, and outputs energy consumption and associated CO₂ emissions by fuel and by end-use (Hughes, 2011). Checks were carried out on outputted energy values to ensure that the data was within the range of the National Energy Efficiency Data-Framework (NEED) values. Energy and CO₂ values are appended to the individual dwelling NCRM record. Values for individual dwellings (electricity and gas) can then be aggregated by any spatial extent or combination of attributes for further spatial analysis.

3.9. NCRF multiple dwelling heating gas consumption estimation

(DECC, 2014a) provides the largest source of data available for analysis of heating gas consumption. The per-property type heating gas consumption only includes properties using gas to heat their homes (i.e. consumption between 2,500kWh and 50,000kWh). Two factors were used to characterise this consumption: (i) the size of a property i.e. floor area; and (ii) number of bedrooms. This paper only covers floor area; however, there is a strong correlation between the two variables (DECC, 2010). It has been hypothesized in the principal components analysis that NEED property size appears to be the strongest individual driver of gas consumption. Also, there is considerable variation in energy consumption for different types of properties and the only physical attribute of a property where the relationship to energy use is less clear is property age (DECC, 2012, p. 2).

In the NEED study (DECC, 2012), four million records (a 17.8 per cent sample) were drawn from the Valuation Office Agency database; this number was selected to allow for attrition and still have a sample of at least ten per cent of properties in England. Four variables were used to ensure the sample was representative; region, property type, number of

bedrooms and age of property. A uniform random sample was drawn from each category. In order to compare the model results by property type, we have used from the NCRM a sample of 50% of the residential dwellings in Newcastle (Calderón et al., 2015, pp. 244 - 246).

3.10. NCRF aggregate-dwelling energy estimates and validation strategies

The cluster model on the MLSOA scale uses six categorical variables (dwelling age, wall construction, building form, dwelling size, heating and number of storeys). The three variables of building form, wall construction and heating are composite variables. The Pearson chi-square tests of association to verify a statistically significant association between two categorical variables (if two variables tend to occur together) so they can be represented by a new categorical variable. According to Pearson Chi-Square the surveyed variables building type and detachment compose the building form variable, while the wall type and wall insulation compose the wall construction variable; finally, the fuel type and heating system compose the heating variable. Each of the six variables was given categorical values according to their influence in the energy demand e.g. lower values to inefficient elements and higher values to efficient elements. The result is a reduced categorical six dimensional data set.

In order to validate the model four sets of observed data were used. The spatial aggregates of 2009 Department of Energy and Climate Change (DECC) consumption data for MLSOA and Lower Layer Super Output Area (LLSOA) were compared against model output spatial aggregate values. In addition, property type aggregates were compared to DECC property type estimates for the North East (DECC, 2013a). A final test was used for Economy 7 aggregates aggregated spatially for MLSOA and LLSOA, as these are treated separately in DECC's consumption figures for MLSOA and LLSOA. A NCRF aggregate estimate of energy consumption was calculated

by summing the individual values for every dwelling across MLSOA and their respective LLSOA, as shown in Equation 5.

$$E_{COM} = \sum_{i=0}^n E_i * N_i \quad (5)$$

Where E_{COM} is the overall predicted annual energy consumption for a community of dwellings (kWh); n is the total number of house archetypes; E_i is the predicted annual energy consumption for house archetype i (kWh); and N_i is the total number of dwellings of house archetype i in the community. The NCRF energy aggregate estimates were validated with 2009 DECC energy estimates.

The annual energy consumption for Castle, South Heaton and Westgate using the cluster model is in Figure 6.

MLSOA	2009 Model Estimation			2009 DECC Estimation			2009 Differences		
	Annual Electricity consumption (kWh)	Annual Gas consumption (kWh)	Annual Total Energy consumption (kWh)	Annual Electricity consumption (kWh)	Annual Gas consumption (kWh)	Annual Total Energy consumption (kWh)	Annual Electricity consumption (%)	Annual Gas consumption (%)	Annual Total Energy consumption (%)
Castle	16,213,837	55,654,385	71,868,222	15,184,682	60,287,546	75,472,228	7%	-8%	-5%
South Heaton	15,267,453	61,412,232	76,679,685	14,228,799	60,075,526	74,304,325	7%	2%	3%
Westgate	22,787,300	72,449,103	95,236,403	27,211,294	47,918,836	75,130,130	-16%	51%	27%

Figure 6: 2009 Annual MLSOA NCRF cluster and DECC estimation

From Figure 6 it can be seen that Castle and South Heaton show a close approximation with the DECC values, whereas Westgate is less close to DECC figures; nevertheless, this is within 27% for the combined electricity and gas annual consumption. One possible reason is that the NCRM database is a collation of diverse data sets concerning building structures and energy information. An explanation of the varying performance of the model within the three case study MLSOAs provides an interesting insight into the importance of locality in modelling energy consumption, even at these scales. Westgate has a much more diverse portfolio of building types (e.g. a high proportion of tall, mixed-use buildings) and energy systems (e.g. district

heating and Economy 7), whereas Heaton and Castle are much more homogeneous in their structures. It may also be a factor that Westgate is a deprived area with a high proportion of social housing and there are interesting socio-demographic issues at play -for example Westgate has two LLSOA areas have the highest percentage of people aged 16-74 in long-term unemployed in the case study areas: LLSOAs 8349 (4.22%) and 8399 (4.88%)- that are not reflected in the modelling process. Although the cluster model performs reasonably well (i.e. at district level) due to the averaging, it does not take into account, local characteristics of the building stock, socio-economic data (e.g. tenure) or surrogates (e.g. the settings of internal temperature, which closely relates to occupancy). This suggests that building level data and comprehensive energy system data is a key component for accurate modelling of energy consumption, even at these large scales.

4. NCRF Model refinement

The refinement process creates a complete data set that improves the quality in the NCRF augmented data. Three key areas of model refinement were noted through empirical testing of the model. The first refinement assigned a better estimate to floor area to houses on multiple occupancies through careful division of the total floor area by the number of residences. The second key refinement manually assigned appropriate heating matches to group buildings where heating is provided through a load sharing mixed residential/commercial existing district heating or group heating. In addition, to a large number of properties (particularly in social housing tenure) in Newcastle that still use Economy 7 heating; and finally, the third key refinement was needed because CHM only uses self-included dwellings. For the third refinement, all registered Houses in Multiple Occupancy (HMO) and shared rented properties were first individualized by council licence or market information, then, to ensure the best match, these Licenses were aggregated and a semi-manual EHS

matching process was carried out.

4.1. NCRM floor area of individual flats in a block and slab

Ordnance Survey MasterMapTM classifies features using feature type and feature description attributes. One of its attribute descriptions is the calculated area value (footprint) of a polygon in square metres, this value usually means the area of the base of the building.

For a block of flats building, reasonable floor area provision for individual flats was attained considering that (i) every individual flat has a floor area no greater than the building footprint, or (ii) the average dwelling floor area for all dwellings in the building is no greater than the building footprint, (iii) the footprint area is the aggregate of the floor area of the individual flats in the floor, and (iv) the total number of residential flats in a building was calculated as the number of residential building floors times the number of flats per floor. The merit of the improvement in the floor area calculation was P_A , the fraction (in percentage) of dwellings affected by better estimates of floor area.

4.2. Cross-validation data sets for improvement on spatial interpolation outcomes

One useful tool to improve any type of spatial interpolation is cross-validation (Efron and Gong, 1983). Cross-validation entails first removing one interpolated observation from the analysis; then attempting a better estimate of the removed value; and finally repeating these steps for every possible estimated data.

In our context cross-validation is the process of removing one anomaly from the previous interpolation and using the survey values available to overwrite the estimation. The Newcastle Council and utilities data sets used

are: (i) District and Group Heating (DGH); (ii) House in Multiple Occupation (HMO); (iii) Economy 7 (E7) fuel type; and, (iv) area correction (A). We propose the use of sparse, informative feature points approach. The use of such features makes the problem of removing anomalies more manageable while providing increased robustness to noise and attribute variation.

In the case of dwellings outside the spatial convex hull area, the improvement procedure involves correcting the spatial interpolation edge effect (SC) by aggregating from neighbour surrounding MLSOAs. The reason for the edge effect is because applying spatial interpolation methods on an individual MLSOA, the method is assumed to lie on the particular MLOSAs convex hull (of the NCRM WarmZone data points inscribed in this MLSOA). The more unvisited dwellings are in the edge, the smaller is the convex hull area. One possible solution to the edge effect is adding the neighbouring surrounding MLSOAs, this increases the convex hull to the NCRM WarmZone centroids of the total MLSOA added, including the original MLSOA. One special case is the south of Westgate because the Tyne River is the neighbour surface; the procedure in this case for solving the edge effect is tedious, it involves a manual search for dwellings with similar profile to be the information donor.

The model refinement approach is described in Equation 6.

$$P_j = f(P_{DGH}, P_{HMO}, P_{E7}, P_{SC}, P_A) \quad (6)$$

Where f represents a particular case study, or in this paper a particular spatial interpolation method, and P_j represents the total fraction of improvements applied to every case study. By comparing improved interpolated values at each dwelling against the total number of dwellings, a direct measure of refinements improvements (P_j) was obtained, allowing the relative merits of each procedure to be assessed.

5. Results

At the smaller level of aggregation of LLSOA, these local area characteristics are expected to show in more detail. Figure 7 displays the DEM model results at the LLSOA level.

	LLSOA	2009 Model Estimation						2009 DECC Estimation						2009 Differences		
		Residential UPRN	Total stock	UPRN No data in class	Annual Electricity consumption (kWh)	Annual Gas consumption (kWh)	Annual Total Energy consumption (kWh)	Electricity + E7 meters	E7/(E7+E)	Annual Electricity consumption (kWh)	Gas meters	Annual Gas consumption (kWh)	Annual Total Energy consumption (kWh)	Annual Electricity consumption (%)	Annual Gas consumption (%)	Annual Total Energy consumption (%)
Castle	8306	729	1,140	33%	2,391,052	10,518,903	12,909,955	734	2%	2,896,451	705	12,512,795	15,409,246	-17%	-16%	-16%
	8307	506	836	27%	1,376,298	5,727,723	7,104,021	490	2%	1,471,463	487	6,419,991	7,891,454	-6%	-11%	-10%
	8308	481	916	35%	1,425,907	5,938,554	7,364,461	611	3%	1,950,176	525	7,995,921	9,946,097	-27%	-26%	-26%
	8311	617	1,015	39%	1,939,494	6,851,613	8,791,107	685	11%	2,694,528	593	9,372,288	12,066,816	-28%	-27%	-27%
	8312	517	731	29%	1,398,933	4,954,461	6,353,394	517	3%	1,928,826	512	7,127,555	9,056,381	-27%	-30%	-30%
	8313	488	630	22%	1,343,307	4,397,440	5,740,747	495	7%	1,833,166	473	6,794,391	8,627,557	-27%	-35%	-33%
	8294	632	929	27%	1,800,304	6,258,031	8,058,335	640	8%	2,410,072	608	9,323,397	11,733,469	-25%	-33%	-31%
South Heaton	8303	1041	1,547	17%	4,377,180	9,699,069	14,076,249	1,094	18%	3,062,105	591	7,326,969	10,389,074	43%	32%	35%
	8359	632	886	25%	2,131,968	10,919,221	13,051,189	630	6%	2,153,830	610	9,644,712	11,798,542	-1%	13%	11%
	8360	853	1,100	20%	3,006,039	12,749,411	15,755,450	854	4%	2,271,434	823	10,675,756	12,947,190	32%	19%	22%
	8361	729	976	21%	2,386,341	11,324,933	13,711,274	739	4%	2,134,881	734	9,860,023	11,994,904	12%	15%	14%
	8362	749	1,043	22%	2,561,513	11,811,979	14,373,492	774	12%	2,286,790	727	10,700,216	12,987,006	12%	10%	11%
Westgate	8364	754	1,171	30%	2,622,484	12,501,002	15,123,486	753	9%	2,319,759	744	11,554,064	13,873,823	13%	8%	9%
	8394	218	519	12%	727,443	2,382,910	3,110,353	222	13%	649,562	251	3,048,655	3,698,217	12%	-22%	-16%
	8395	840	1,309	28%	2,750,815	11,714,160	14,464,975	751	1%	1,739,168	636	7,989,220	9,728,388	58%	47%	49%
	8397	687	2,185	31%	2,444,095	8,127,934	10,572,029	684	76%	3,667,679	288	5,835,231	9,502,910	-33%	39%	11%
	8399	818	1,220	31%	5,176,186	18,152,000	23,328,186	800	11%	2,349,583	509	6,329,353	8,678,936	120%	187%	169%
	8439	849	1,086	16%	2,960,014	10,596,295	13,556,309	868	13%	2,556,256	718	8,661,471	11,217,727	16%	22%	21%
	8440	2580	6,593	25%	10,052,576	37,779,137	47,831,713	2,801	41%	16,249,046	863	15,848,757	32,097,803	-38%	138%	49%

Figure 7: 2009 Annual LLSOA NCRF sub-city DEM and DECC estimation

From Figure 7, there is a consistent underestimation of NCRF combined energy results in Castle, and an overestimation in South Heaton and Westgate. These results are affected by both model and data assumptions and issues with the DECC data. In the results presented in Figure 6, there is a consistent underestimation of the energy model in all LLSOAs in Castle.

Figure 8 shows the mean and median for the aggregated sample (Castle, South Heaton and Westgate) in four different property types.

Figure 8 shows that the mean and median NCRF annual gas consumption is higher than the NEED values. Because the results correspond to a model which is driven by a data set that ultimately corresponds to individual houses with known address, one can conclude that local area characteristics play an important part in the energy consumption (e.g. most of the pre-1919 mid-terraced and end-terraced

corresponds to uninsulated solid wall terraced properties with higher energy consumption, and the 1965 to 1982 cavity terraces are also mostly uninsulated in Castle, South Heaton and Westgate).

Sample		Property Type	Property Age	Floor Area (sqm)	NCRF			NEED			Mean Difference (%)	Median Difference (%)	Wilcoxon Signed Ranks		
					Sample Size	Mean Annual model estimated heat gas consumption (kWh)	Median Annual model estimated heat gas consumption (kWh)	Sample Size	Mean Annual DECC estimated gas consumption (kWh)	Median Annual DECC estimated gas consumption (kWh)			Negative ranks	Positive ranks	Ties
1	House	End-terraced	Pre 1919	50 to 75	57	16,343	13,919	150	11,483	10,905	42%	28%	5	52	0
2	House	End-terraced	1965-1982	50 to 75	82	16,185	17,901	180	9,418	9,252	72%	93%	2	80	0
3	House	Mid-terraced	1965-1982	50 to 75	147	14,821	14,767	340	8,509	8,178	74%	81%	21	127	0
4	House	Mid-terraced	Pre 1919	50 to 75	486	14,746	13,919	1,380	9,500	8,757	55%	59%	23	467	0

Figure 8: NCRF and NEED data sets for heating gas consumption by property type

At this level, retrofit decisions can be made, whether to improve the efficiency of the supply side technologies, or invest in demand side technologies with the intention of reducing primary energy requirements. Residential building retrofits are at the forefront of the sustainable development agenda if government building regulation is taken as a proxy.

6. Discussion

Section 5 describes the use of the Cambridge Housing Model (CHM) to estimate the annual energy consumption of a single dwelling, and develops an aggregation strategy for small areas within the city (sub-city areas). This section discusses the framework for estimating the modelled energy results in terms of the modelling methods and the spatial data used to validate the results for a single dwelling as well as sub-city areas. The NCRF framework is comprised of data sets, energy modelling aggregation and validation methods. This section discusses the assumptions within the developed framework and how these affect the analysis of the results.

Four classes of data sets were used to build NCRM: dwelling domestic stock data, building physics data, household data and climate data. The assumptions in the modelling methods are described in Sections 6.1 to

6.3 and DECC and NEED data is used as a validating data set is in Section 6.4. The NCRM has two data sets at the resolution of the individual dwelling, with one data set of rough approximations of household occupancy and three average regional scale landscape and climatic data sets. The NCRM individualized data sets are the dwellings domestic stock data and building physics data. The coverage of both data sets is different; whilst the domestic stock data have 100% coverage, the buildings physical data do not. The spatial interpolation strategies are used to complete the missing survey data, described in section 3, Section 3.6.

6.1. The Cambridge Housing Model (CHM)

The CHM is the basis for the per-dwelling energy estimates. However, there are a number of assumptions built into the model that potentially impact on the results of the Energy Modelling Framework. The main assumptions in using the CHM in the framework concern the use of averages for climate data and occupancy.

The CHM relies on regional and monthly climate data as part of the energy use calculations. Additionally, energy outputs are adjusted to match the national statistics in the Digest of UK energy statistics (DUKES) for gas and electricity use since 2009 (Hughes, 2011, p. 14). This means the same weather adjustment in the monthly external temperature ($^{\circ}\text{C}$), monthly average wind speed (m/s) and monthly average horizontal solar radiation (W/m^2) are in place for every dwelling in the North East of England, disregarding any possible local microclimates.

Urban areas tend to have higher air temperatures than their rural surroundings, as a result of gradual surface modifications that include replacing the natural vegetation with buildings and roads. This is because vegetation plays a significant role in regulating the urban microclimate and

can influence domestic energy demand through solar absorption and the cooling effects provided by shade and evapotranspiration (Akbari et al. (2001); Akbari and Konopacki (2005)). This may mean that areas with a low residential density indicative of more open space require more energy to maintain the same temperature as higher density areas.

Additionally, CHM uses a standard occupancy model, which means a standard number of occupants, i.e. the number of occupants is related to the number of rooms and hence the floor area, and also to standard heating regimes, i.e. the zone one heating times are applied throughout the dwelling for nine hours during weekdays and for 16 hours at the weekend.

Assessing household energy requirements using standard occupancy enables comparison of the energy performance of similar dwellings across Newcastle's housing stock. However, standard occupancy does not provide an understanding of the way a household actually uses energy within the dwelling. Actual energy consumption is determined by a wide range of factors beyond simply the physical characteristics of the property. These include socio-demographic, economic, behavioural and practice factors which all influence the amount of energy consumed within a household (Druckman and Jackson, 2008; Andersen et al., 2009; Abrahamse and Steg, 2011).

On the whole, the majority of households tend to consume less energy than is estimated by the SAP calculation, especially low energy buildings (Kelly et al., 2012, p. 6864). In thermal terms, the worse a home is, the more economically the occupants tend to behave with respect to their space heating; this phenomenon is labelled the pre-bound effect (Sunikka-Blank and Galvin, 2012, p. 1270).

In summary, these assumptions in the CHM may be manifest in the energy modelling framework in areas of low residential density, where the energy consumption may be underestimated due to microclimate effects and where

there are distinct local idiosyncrasies in the socio-demographic make-up of an area that is divergent from the standard occupancy.

6.2. Creating a full SAP record

The process of creating a full SAP record for use in the CHM is done through record augmentation in a two stage process. The first involves making the best match between a complete NCRM record and an EHS record, and then copying over unknown fields in the NCRM data set from corresponding EHS fields. The second process involves extending these results to related properties to provide a suitable SAP record for every dwelling.

Spatial interpolation introduces uncertainties. The two which are related are the choice of interpolation method and the treatment given to the data outside 'edge effect,' i.e. in order to fully cover the surface of an interpolated area, some unknown points around the edges of the data set needs to be extrapolated. The third is in the urban penetration of the physical survey.

- i. In order to create a complete data set, spatial relationships between dwellings must be inferred. To do this a variety of spatial interpolation methods depending upon the physical characteristics of the urban form were used. The choice of a SA model differs in the way it infers interpolation surfaces: in Castle and South Heaton, surfaces are inferred from deterministic variables (clusters in Castle and density in South Heaton) or inferred from stochastic variables (distance in Westgate). In addition, all the interpolation methods suffer to some degree from edge effects and data does not fit inside the interpolated surface. Depending on the availability of other data (e.g. from neighbouring MLSOAs), it was possible to extend the interpolation but in some areas this had to be done manually due to physical barriers

(such as the River Tyne). It is likely that the choice of SA method may have some, if relatively minor, effect on the final energy aggregation estimates.

- ii. Perhaps of more importance is the spatial heterogeneity of the survey data upon which the initial match between NCRM and EHS data is made. This is particularly important where there are local energy structures such as E7, HMO and District and group heating. This is a major concern in Westgate in particular where most of these structures occur and the physical survey penetration is relatively low. To partially overcome this issue, secondary information from the local housing markets was brought into the analysis.

Record Augmentation uses multiple imputations of EHS variables applied to a NCRM individual record to produce an individual, extended NCRM record with full SAP information. The fact that a perfect imputation leading to a single full augmentation with all ten base variables from the initial NCRM record is only possible on 11% in South Heaton and 1% in Westgate makes imputation modelling potentially more relevant than interpolation. There are several reasons why perfect imputation is not possible:

- i. The EHS and the Warm Zone (the physical data of NCRM) programme studies have different purposes, universes, sample sizes and accuracy.

The purpose, universe and sample size of the studies are different. The EHS is a detailed survey of England's household and dwelling characteristics to inform the development and monitoring of DCLGs housing policies, whereas the overall aim of the Warm Zone programme purpose was to facilitate the delivery of practical domestic energy efficiency measures to alleviate fuel poverty at a spatial resolution below that of the Local Authority.

The national EHS sample is comprised of 16,150 occupied or vacant dwellings where a physical inspection was carried out and includes 15,512 cases where an interview with the household was also secured. These are referred to as the dwelling sample and the household sub-sample respectively; household results are then weighted to population totals by age, sex and region, and to the tenure distribution, whereas the Warm Zone programme comprises 60,977 detailed energy property surveys conducted by a visit to the property by a Warm Zone representative and an informal interview with the householder.

The accuracy in terms of the percentage of the entire population examined in both studies is different. EHS covers 0.073% of the English housing stock and the Warm Zone programme covers approximately 50% of the Newcastle residential housing stock. While EHS uses a representative sample, NCRM WarmZone has a bias toward households living in dwellings that could be potentially fuel poverty beneficiaries'. This is further discussed in Section 6.5 under the topic of contributing forms of uncertainty.

- ii. Multiple NCRM records will select" the same EHS record. Often there is more than one NCRM record that matches a single EHS record; this makes the imputation method relevant. The assumption is to use the nearest neighbour imputation (NNI) method, in which the criteria is to determine which EHS record is the most likely in accordance with the fusion variables, the principal components and factor analysis. The closest EHS record is then used as the donor.
- iii. Some dwellings do not fit the EHS data. The EHS dwelling records are all self-contained units of accommodations (DCLG, 2009, p. 13), whereas NCRM include additionally non-self-included accommodations and HMOs. In Newcastle, the term HMO refers to a rental property which falls into one of the following criteria: a house which is split into bedsits; a house, or flat-

share, where each of your tenants has their own tenancy agreement; or students who live in shared accommodation. The assumption is first to consider only those HMO licences issued on (or before) the year of the physical survey, and second the number of rooms providing sleeping accommodation as a rough approximation of the number of rooms of a synthetic self-containing dwelling. South Heaton and Westgate have a significant number of HMOs mainly due to student shared accommodations.

- iv. Attributes do not align perfectly across the data sets. Two fields from NCRM did not align with EHS corresponding field definitions: detachment and dwelling age. 4.2% of Newcastle's housing stock is made up of enclosed detachment dwellings from back-to-back houses and high rise buildings. This form of property type is not considered in the EHS survey. The assumption in NCRM is to consider enclosed (three shared walls) properties as a regular mid-terrace (two shared walls).

The age classifications based on the NCRM data (derived mainly from NCRM Cities Revealed data) did not match that of the EHS dwelling age. Proportional mapping (the proportion of the EHS age band that lies within each of the corresponding two NCRM Cities Revealed band) is used to map EHS age band to NCRM age bands where necessary.

YHN own approximately 30,000 properties throughout the city, meaning an accurate build date can be taken from the deeds. Therefore any mismatch with other data sets (e.g. NCRM Cities Revealed) can be easily corrected.

In summary, creating a complete full data set of extended NCRM records required a considerable number of refinements in South Heaton and Westgate, where the Warm Zone programme has lower penetration and complex infrastructures are in place.

6.3. *The number of thermal zones*

The number of thermal zones in a building is determined by various factors including building size, shape, orientation, and usage, whereas CHM, BREDEM and SAP models use a simplified model for the number of thermal zones. The first zone comprises the major occupied spaces (living area) of a dwelling, while the second zone is the bedroom and the rest of the house. The current assumption in SAP is that the heating thermostat is set to 21°C and 18°C in the two zones respectively; however, a default demand temperature of 19°C is assumed in CHM for the living area and for all dwellings, as opposed to 21°C in SAP, independent of building massing.

Building massing is one of the most important factors in passive solar design, and is one of the most important stages in the design of a building, as the surface areas are exposed to sun at different times of day. The building height and width can all be optimized for passive comfort.

This assumption significantly reduces the amount of time required for energy modelling by simplifying the internal layout of the building. In CHM, the internal spaces of every dwelling in a linear block of terrace houses and tall buildings are divided into zones, and there is no difference in their usage and activity, heat gains and losses characteristics, heating systems and operational regimes.

Internal layout or thermal zoning has not been considered as an uncertainty variable in other domestic energy models which focus on energy performance assessment using steady-state models (Firth et al., 2010). CHM Hughes et al. (2013) does not consider the number of thermal zones as one of the 31 most sensitive parameters in the local sensitivity study. However, (Korolija and Zhang, 2013, p. 1190) found the impact of this model simplification to have a mean (absolute) error across diverse buildings of 10.6% for annual heating demand.

In the NCRM/CHM modelling described in Section 3.9, a linear block of terrace houses ignores the heat flow from one terrace to the adjacent terrace, i.e. it was relatively safe to assume that the temperatures in the houses on either side were be roughly the same as that in the modelled house, so the thermal transfer through these side walls was likely to be negligible compared to the losses and gains from the outside environment. This means that there is no net flow of heat in either direction, or the dividing walls between two adjacent terraces were adiabatic. A similar condition occurs in the model of a single floor in the middle of a high-rise building.

In the UK, the Dwelling Fabric Energy Efficiency (DFEE) methodology was adopted within the Code for Sustainable Homes (November 2010 version). DFEE methodology considers the space heating demand of a dwelling as being affected by: building fabric U-values, thermal bridging, air permeability, thermal mass, and features which affect lightning and solar gains (Zero Carbon Hub and AECOM (2012)). In particular for buildings containing multiple dwellings, AECOM showed the air permeability of the full DFEEs drop from 5.2 ($\text{m}^3/\text{hr}/\text{m}^2$ @ 50Pa) for a three dwelling terrace to 4.8 ($\text{m}^3/\text{hr}/\text{m}^2$ @ 50Pa) for a seven dwelling terrace.

AECOM suggests that air permeability affects the space heating demand of mid- terrace dwellings such as the ones in South Heaton, where there is a high concentration of multiple dwelling terrace houses.

In summary, the BREDEM model treats dwellings as individual entities and does not take into account adjacent dwellings with shared party walls and floors. The potential heat gain from these buildings is not estimated in the BREDEM model and hence NCRF may overestimate energy in building types like linear terraces where this effect may be measurable. In reality, there is an increased heat loss due to air leakage through exterior envelopes as well as the interior partition walls, ceilings and floors (Kusuda, 1977, p. 97), therefore,

heating energy is needed to offset the heat loss (Lindberg et al., 2008). The number of thermal zones will vary depending on many factors including the building use, size, and shape, where a 'zone' is a segment of a building with similar thermal requirements serviced by the same mechanical equipment and controls. On the number of thermal zones, the US DOE (2013, p. 34) EnergyPlus™ simulation program indicates that "the minimum number of zones in a general simulation model will usually be equal to the number of systems serving the building. The collection of heat transfer and heat storage surfaces defined within each zone will include all surfaces bounding or inside of the space conditioned by the system."

6.4. DECC and NEED data as a validating data set

Another potential source of difference between DECC consumption aggregates and NCRF energy output may be attributed to underlying differences or potential inaccuracy within the DECC aggregate data. The main discrepancies are electricity and gas meter count (DECC, 2008), properties not being connected to the gas grid, the DECC domestic/industry threshold, bulk power agreements, and weather correction methodology.

- i. To produce the electricity consumption aggregated estimates, annualised consumption data were provided to DECC at Meter Point Administration Number (MPAN) level by the data aggregators (DAs). DAs are agents of the electricity suppliers, who collate/aggregate electricity consumption levels from each electricity meter.

MPAN consists of approximately 80 per cent actual ("Annual Advance") readings and 20 per cent estimated readings ("Annual Consumption"). In addition, electricity consumption data for each MPAN is not weather corrected. Furthermore, the sum of meter points for domestic energy consumption at MLSOA level does not always equal the sum of meter

points for domestic energy consumption at the associated levels due to missing or incorrect address information or an unallocated load.

- ii. DECC (2013) identify areas and types of properties off the gas grid. It shows that purpose-built flats are the least likely to have a gas meter. The NEED data also suggest that, small modern properties are the least likely to have a gas connection, for example, 70 per cent of post 1999 flats have no gas connection. It is likely the majority of these properties without gas are in areas which are on the gas grid, but with no gas connection in the property (e.g. blocks of flats in high density urban areas). The data also show that more generally, the largest and smallest properties are least likely to have a gas connection, for example 43 per cent of properties with a floor area of more than 200 square metres and 42 per cent of properties with a floor area of 50 square metres or less have no mains gas in the home; compared with the average for all properties of 18 per cent (DECC, 2013, p. 74). Properties in Westgate LLSOAs 8397 (58.08%), 8440 (66.55%) and 8349 (37.78%) are off the gas grid properties, and are therefore not included in the DECC gas statistics but through interpolation are included in the NCRM data set.
- iii. DECC sub-national statistics use an industry cut off of 73,200kWh to determine whether a gas meter is domestic or not, with all meters with consumption of 73,200kWh or below assumed to be domestic. This means a number of smaller commercial/industrial consumers are allocated as domestic and therefore the estimates of the percentages of households without gas is an underestimate of the true number.
- iv. Your Homes Newcastle (YHN) tenants are billed directly for energy use, except in properties supplied by district and group heating networks (charged by a fixed tariff). YHN and Newcastle City Council purchase the fuel for district and group heating in bulk, together with the other councils in the North East. DECC statistics (which reports individual meter readings) do not include these properties.

- v. DECC gas sub-national consumption figures have been weather corrected, whereas electricity consumption figures are estimates of actual consumption and have not been weather corrected (DECC, 2014b, p. 10). The DECC annualised and weather corrected Meter Point Reference Number (MPRN-level) gas consumption data are obtained from XoServe. The National Grid apply the weather correction to the data prior to it being supplied to XoServe and DECC.

Although a certain amount of information relating to the DECC process of annualisation and weather correction is available (National Grid, 2012, p. 28), the effect that this process has on modifying the gas values, and therefore the impact that the factoring has on the final results of analysis, is not fully known. This process of weather correction is therefore impossible to reverse based on information currently in the public domain and introduces unknown effects into the available annualised records for each individual house that cannot be interrogated (Hamilton et al., 2011). In the case of NEED gas consumption data, these are weather-corrected by the Energy Companies before being submitted as annual data to DECC.

NCRF uses CHM for estimating the annualised gas consumption. CHM Outputs are also adjusted to match DUKES data for gas and electricity use each year. This means that effectively it also includes a weather-adjustment (Hughes et al., 2013, p. 3). The key modification in CHM is the use of 19°C (292°K) as the baseline demand temperature for the living area for all dwellings SAP uses 21°C.

In CHM, a normalized sensitivity coefficient (S_i) is computed to consider the impact on the total energy consumption of small changes to internal demand temperature model parameters x_i , as shown in equation 7.

$$S_i = \frac{dy}{dx_i} \frac{x_i}{y} \quad (7)$$

Where (xi) represents the original value of the internal demand temperature (°K) model parameter, (y) represents the original total energy consumption, and (Si) represents the sensitivity of the model output to parameter variations around the best estimate values. The normalized Si equals 23.31 (Hughes et al., 2013, p. 159) for variations in the internal demand temperature individually, while holding constant the others parameters affecting the energy demand.

By choosing dxi to represent a 1% variation in xi, Si represents the percentage change in total energy due to a 1% change in the model parameter (e.g. the original demand temperature is 292.15°K or 19°C; a 1% variation expressed in Kelvin equates to a variation of 2.92°K, while for °C it equates to 0.19°C). A variation of 5.2632% expressed in °C equates to a variation of 1°C and the value for dy/y is then equal to 1.2268. In Castle, the results show a change in the demand for gas to 99,419,327kWh (from an original value of 44,646,725kWh), and for electricity to 25,998,546kWh (from an original value of 11,675,295kWh). The internal demand temperature is by far the most significant parameter in the CHM modelling process, i.e. for a 1°C increase in the internal demand of temperature; there is an increase of 122.68% in the energy consumption.

In summary, different weather correction methodologies in DECC/NEED and CHM lead to some discrepancies, which probably vary somewhat from year to year. The main reason for this is that the exact methodology for weather correction for NEED (and DECC) is not fully disclosed.

This section has discusses some of the key assumptions that need to be considered when comparing NCRF energy outputs with DECC and NEED

data. The actual effect of any individual issue is difficult to quantify but detailed analysis may provide some areas for further discussion.

7. Conclusions

This paper has described an annual energy consumption model for sub-city areas in UK cities, with Newcastle upon Tyne as a case study. The energy modelling approach is top-down (a cluster approach) in districts and bottom-up (a sub-city DEM approach) in neighbourhoods and communities. Three districts were studied in detail, Castle, South Heaton and Westgate, along with the corresponding nineteen neighbourhoods. The annual energy consumption (electricity and gas) was estimated using the year 2009 as a base scenario. The spatially enabled input database included four main data sets: two local data sets (NCRM WarmZone and NCRM Gazetteer) and two national data sets (English Housing Survey and Ordnance Survey). All data sets used are generally available in some form to many if not all local authorities. The energy model used to estimate the energy end-use was the Cambridge Housing Model (CHM). CHM is a national model requiring a full SAP input (115 variables). The suitability and plausibility of the model was verified against comparable DECC statistics.

Validating (with respect to DECC consumption values) the modelled data at MLSOA/LLSOA geography requires full coverage within the whole district. However, the data do not have full coverage in these geographies nor do they have the complete SAP input records. To fulfil these requirements, this paper has proven that record generation (for missing records to a complete coverage) and augmentation (for missing fields in every record) algorithms lead to the desired full coverage SAP input in the MLSOA/LLSOA. Spatial interpolation algorithms were used as a record generation strategy and a nearest neighbour multiple imputations algorithm was used as a record augmentation strategy. Both had reasonable success, given that performing a

physical survey for full SAP data on each and every dwelling in Newcastle is not possible.

A new spatially enabled framework approach for modelling energy consumption in sub-city areas is proposed. Previous modelling approaches adopted a number of house archetypes which together represent all dwellings in the stock. This paper identifies individual dwellings energy profiles to estimate individual and aggregated energy consumption. The cluster model uses six categorical variables to capture the mean effects in the interaction between the physical variables in districts as the individual effect will cancel out at this scale. The framework estimates in the sub-city DEM model are improved in neighbourhoods by taking into consideration ten variables and local micro-cohesive structures. The reason for the increased number of variables is to obtain a better insight into the effect of uncertainties of the aggregated process by generating a more complete representation of the dwelling. The framework estimates the energy consumption in individual properties and explores the potential use of the NCRF in retrofit campaigns.

The framework is likely to be of significant value as a policy tool. It can be used to inform and direct policy by testing the effect that various policy decisions are likely to have on the community energy consumption. For instance, the framework could be used to test the effect of possible future revisions to the Building Regulations or energy efficiency schemes. The framework will significantly increase our theoretical understanding of the complex inter-relationships that exist, not only between the various end-uses of energy within the demand side of the European housing stock, but also the relationships that exist between the demand side and a particular combination of urban form, climate, buildings physical characteristics and behaviour and household social practice in a sub-city area.

Difficulties in the validation point to further studies on: (i) thermal zones;

(ii) scale of analysis, and (iii) urban form. In the thermal zones, the internal layout or zoning has not been considered as a variable on domestic energy models focused on energy performance assessment using steady-state models, in which by now the internal zones are limited to two. Evidence shows that in linear terraces a multi thermal zone could give better results but at a cost of increased model complexity. In the scale of analysis: A cluster model appears to be satisfactory at MLSOA and cities scale, where neighbourhood effects on energy consumption cancel out, but do not disappear, although in a neighbourhood scale a sub-city DEM model that takes into consideration cohesive structures potentially gives more sensible results. Finally, in the urban form issue, the evidence for the influence of urban configuration (form) and outdoor climate conditions, and finally on the energy balance of buildings, is quite conclusive; an accurate modelling of the interface between an urban surface and the atmosphere above is needed, but the exact characterization of this issue is complex.

The validation also shows that there are a significant number of uncertainties which are not usually communicated and understood by LAs, policy development of national data sets, and other stakeholders such as planners, architects and engineers. To better understand, communicate and describe uncertainties, it is necessary to obtain a detailed local knowledge of the stock and non-filtered building or post code level consumption data. Further model validation to ascertain uncertainties in the CHM model, imputation algorithms, and DECC data is certainly needed. In this study, we go even further to suggest that a re-think of underlying energy models to enable the integration of building and urban modelling challenges are necessary.

8. Appendix

A. Descriptive Statistics of the NCRM sample

Dwelling built form			
		Frequency	Per cent
Individual records	Bungalow	6,420	8.2
	Flat	19,490	24.8
	House	51,802	66.0
	Maisonette	763	1.0
Total		78,475	100.0

Table A.1 NCRM WarmZone dwelling built form

Dwelling detachment			
		Frequency	Per cent
Individual records	User-missing (N/A)	59	0.1
	Detached	5,087	6.5
	Enclosed	3,315	4.2
	End-Terraced	10,740	13.7
	Mid-Terraced	25,290	32.2
	Semi-Detached	33,984	43.3
Total		78,475	100.0

Table A.2 NCRM WarmZone dwelling detachment

Dwelling age band			
		Frequency	Per cent
Individual records	1900-1918	7,111	9.1
	1919-1944	23,961	30.5
	1945-1964	19,541	24.9
	1965-1975	13,269	16.9
	1976-1980	4,903	6.2
	1981-1990	3,592	4.6
	1991-1995	1,759	2.2
	1996-1997	966	1.2
	1998-2002	24	0.0
	Post 2002	42	0.1
	Pre-1900	3,307	4.2
Total		78,475	100.0

Table A.3 NCRM WarmZone dwelling age band

Dwelling number of storeys			
		Frequency	Per cent
Individual records	1	25,786	32.9
	2	49,965	63.7
	3	2,404	3.1
	4	66	0.1
	5	15	0.0
	6	5	0.0
	7	1	0.0
	9	1	0.0
Total		78,243	99.7
System-missing		232	0.3
Total		78,475	100.0

Table A.4 NCRM WarmZone dwelling number of storeys

Dwelling wall type			
		Frequency	Per cent
Individual records	User-missing (N/A)	1,390	1.8
	Cavity	64,704	82.5
	Solid	11,613	14.8
	Timber	768	1.0
Total		78,475	100.0

Table A.5 NCRM WarmZone dwelling wall type

Dwelling wall insulation			
		Frequency	Per cent
Individual records	User missing (N/A)	2,012	2.6
	Don't know	13,788	17.6
	Insulated	22,786	29.0
	Uninsulated	39,889	50.8
Total		78,475	100.0

Table A.6 NCRM WarmZone dwelling wall insulation

Dwelling tenure type			
		Frequency	Per cent
Individual records	Housing Association	3,169	4.0
	Local Authority	22,648	28.9
	Owner Occupied	44,008	56.1
	Private Rented	8,650	11.0
Total		78,475	100.0

Table A.7 NCRM WarmZone dwelling tenure type

Heat Fuel			
		Frequency	Per cent
Individual records	User missing (N/A)	659	0.8
	Coal	100	0.1
	Communal	122	0.2
	Economy 7	5,436	6.9
	Electric	360	0.5
	Gas	71,710	91.4
	Liquid Petroleum Gas (LPG)	8	0.0
	Oil	77	0.1
	Wood	3	0.0
Total		78,475	100.0

Table A.8 NCRM WarmZone heat fuel

Heating System			
		Frequency	Per cent
Individual records	User missing (N/A)	420	0.5
	Combinational boiler	37,139	47.3
	Communal	1,885	2.4
	Condensing boiler	1,394	1.8
	Condensing combinational boiler	1,623	2.1
	Room heater	1,090	1.4
	Standard boiler	28,394	36.2
	Storage heater	4,640	5.9
	Warm air	1,890	2.4
Total			100.0

Table A.9 NCRM WarmZone heating system

B. The local data set: NCRM level envelope parameters and tenure

This Appendix provides the descriptive statistics for the NCRM WarmZone level envelope parameters and tenure system variables.

The 'SAP profile' data set is based on an integration of primary and secondary per- dwelling spatial information. The integration is on the following city themes: location; building form; parameters that affect the thermal insulation of the building envelope as area; windows insulation; heating fuel; boiler type and others. Reader may refer to (Calderón et al., 2012) for additional information on NCRM initial data sets. The NCRM utilises a Unique Property Reference Number, a part of the National Land and Property Gazetteer to spatially reference building level data. Additionally, the Unique Property Reference Number is used to link existing and new data sets to augment the existing data on different scales. The resulting data set has 122,733 records of dwellings of which there are 60,977 complete records (using an extended set of energy attributes) and 78,475 records that contain the basic building profile (built form and age band) and some of the extended attributes. In the extended attributes an individual record may be missing, interestingly, there are 1,390 records of the 78,475 where wall type was not recorded. The basic local data set is the NCRM WarmZone, Warm Zone is a not for profit company that administers schemes for providing energy audits for domestic properties and grants for interventions in the UK. NCRM Newcastle WarmZone statistics shows that:

- 24.8% of the dwellings are flats, and 74.2% are houses. Bungalows are 8.2% and Maisonettes are 1%. For a full breakdown of NCRM WarmZone dwelling built form count see Appendix A Table A1.
- Combined detached are 49.8%, the detached group is 6.5% whereas the semi- detached group is 43.3%. Combined terraced dwellings are 50.1%,

the mid-terraced group is 32.2%, the end-terraced group is 13.7% and the enclosed group is 4.2%. Finally, there are 0.1% of the dwellings whose detachment is defined by the Warm Zone surveyor as not available (N/A) (i.e. it is missing). For a full breakdown of NCRM WarmZone dwelling detachment count see Appendix A Table A2.

- Overall, 9.1% of homes were built before 1919. Almost a similar proportion of dwellings were built after 1990 (8.1%). For the other dwelling ages the count is: for the range 1919-1944 is 30.5% and for the accumulated range 1945-1980, the number of dwellings is 48%. For a full breakdown of NCRM WarmZone dwelling age band count see Appendix A Table A3.
- The biggest count of dwellings is the two storeys attribute (63.7%) whereas 32.9% were one storey dwellings, and a little more than 3% of the dwellings have three or more storeys. Also, 0.3% of the dwellings were blanks in the data set. For a full breakdown of NCRM WarmZone dwelling number of storeys count see Appendix A Table A4.
- The majority of dwellings are cavity (82.5%) wall type, solid walls are 14.8%. Also, there are 1.8% (wall type) and 2.6% (wall insulation) of the dwellings whose detachment is defined by the Warm Zone surveyor as not available (N/A). For a full breakdown of NCRM WarmZone dwelling wall type and wall insulation count see Appendix A Table A5 and Table A6 respectively.
- The biggest count of dwelling is the 56.1% owner occupied and the rest were split between the private rented sector (11%) and social rented sector (32.9%). There were considerable less housing association dwellings (4%) than local authority dwellings (28.9%). For a full breakdown of NCRM WarmZone dwelling tenure type count see Appendix A Table A7.

In summary the descriptive statistics in this section helps to describe data in a meaningful way such that patterns might emerge from the data. However, an area- based spatial approach allows more houses to be targeted in places where local area characteristics show inefficient

elements, and may therefore potentially capture a greater number of e.g. fuel poor households per unit of cost, compared to the existing self-referral method, e.g. Green Deal (the UK Green Deal scheme provides finance to make energy-saving improvements in a home and finds the best way to pay for them). NCRM statistics suggests that a detailed data set that incorporates local area characteristics at the level of the property itself and an energy model that can be applied to each building has the potential to provide better information for understanding the consumption patterns and particularly the retrofit opportunities at local scales. For instance, refurbishing a group of Victorian houses with external solid wall insulation thus has several benefits. Yates (2006) argues that the refurbishment of solid walled houses can achieve SAP ratings equal to or better than those of new build properties complying with building regulations by adopting such measures as: 'loft insulation, insulated dry lining to external walls, or external insulating render, ground floor insulation, secondary glazing, gas central heating with condensing boiler, controlled ventilation system', and this may increase the rental income among other benefits not related to energy efficiency.

C. The local data set: NCRM level fuel supply for heating systems

This Appendix provides the descriptive statistics for the NCRM WarmZone heating fuel and heating systems variables.

i. Heating Fuel

Most of the Newcastle dwellings are using gas-mains (91.4%). Economy 7 (E7) dwellings are 6.9%. Other attributes for heat fuel with an overall frequency of less than 1.7% are not representative in the city. For a full breakdown of NCRM WarmZone dwelling heating fuel count see Appendix A Table A8.

ii. Heating system

There are three methods of space heating which are defined within the context of NCRM: (i) central Heating, (ii) storage heaters and (iii) room heaters. See Appendix A Table A9 for the descriptive statistics of the heating systems.

Central heating systems are systems able to distribute heat to at least one room other than that which contains the boiler. Combination (combi) boilers are 47.3%, condensing boilers are in 1.8%, standard boilers are 36.2%, and condensing combi-national boilers are in 2.1%.

Condensing boilers are 3.9% in the domestic homes. This type of boiler improves the efficiency to over 90% in many cases as opposed to around 78% efficiency for a combination boiler. All boilers installed since March 2005 have to be of the condensing variety. Some of these units can be coupled with an indirect fired water heater to provide for the dwelling's hot water needs, and are called condensing combined boilers and represent 2.1% of the stock.

Storage heaters dwellings are 5.9% whereas room heater dwellings are 1.4%. Storage Heaters are the choice for homes that do not have a boiler and radiators. For a full breakdown of NCRM WarmZone dwelling heating fuel count see Appendix A Table A9.

In summary, this section has used descriptive statistics to identify dwellings that have specific interaction with gas consumption. The statistics throughout this section re-iterate the key messages of this paper, which is that local area characteristics are important in understanding the energy consumption estimates in the sub-city areas. As an example, for a group of rented and owner occupied tenures the Warm Front scheme (WF) provides grants for packages of insulation and heating improvements including central heating systems.

D. The National EHS data set

EHS is a national survey of housing in England, commissioned by the Department for Communities and Local Government. The 2009 EHS summary of tables for the key measures of stock and household data is published on government sites (UK Data Service, 2012) alongside a report (DCLG, 2010). The EHS provides the underpinning data for the NCRM individual building modelling strategy. This paper shows in Section 7 that by combining data from different geographies and sources it is possible to augment limited NCRM WarmZone data with EHS data to create a full NCRM SAP input for use in energy models.

The EHS sample comprises of 16,150 occupied or vacant dwellings where a physical inspection was carried out. This is referred to as the 'dwelling sample'. There are 15,512 cases where as well as a physical inspection, an interview with the household was also secured. This paper put the EHS files through a rigorous editing so it can be integrated to the NCRM sample using imputation techniques. The next section describes the methodology of cross-study analysis to integrate the NCRM and EHS.

9. References

- Abrahamse, W. and Steg, L. (2011) 'Factors Related to Household Energy Use and Intention to Reduce It: The Role of Psychological and Socio-Demographic Variables', *Research in Human Ecology*, 18(1), pp. 30-40.
- AECOM (2012) *Fabric Energy Efficiency for Part L 2013*. [Online]. Available at: http://www.zerocarbonhub.org/sites/default/files/resources/reports/Fabric_Standards_for_2013-Worked_Examples_and_Fabric_Specification.pdf.
- Akbari, H. and Konopacki, S. (2005) 'Calculating energy-saving potentials of heat-island reduction strategies', *Energy Policy*, 33(6), pp. 721-756.
- Akbari, H., Pomerantz, M. and Taha, H. (2001) 'Cool surfaces and shade trees to reduce energy use and improve air quality in urban areas', *Solar Energy*, 70(3), pp. 295-310.
- Andersen, R.V., Toftum, J., Andersen, K.K. and Olesen, B.W. (2009) 'Survey of occupant behaviour and control of indoor environment in Danish dwellings', *Energy and Buildings*, 41(1), pp. 11-16.
- Anderson, B. and Chapman, P.F. (2010) *Bredem-12 Model Description: 2001 Update*. London: BRE.
- Ascione, F., De Masi, R.F., de Rossi, F., Fistola, R., Sasso, M. and Vanoli, G.P. (2013) 'Analysis and diagnosis of the energy performance of buildings and districts: Methodology, validation and development of Urban Energy Maps', *Cities*, 35(Supplement C), pp. 270-283.
- Bleiholder, J. and Naumann, F. (2009) 'Data fusion', *ACM Comput. Surv.*, 41(1), pp. 1-41.
- BRE (2009) *The Government's Standard Assessment Procedure for Energy Rating of Dwellings 2005 edition, revision 3*. Garston, Watford: BRE. [Online]. Available at: <http://projects.bre.co.uk/sap2005/>.
- Burns, R.P. and Burns, R. (2008) 'Cluster Analysis', in *Business Research Methods and Statistics Using SPSS*. London: SAGE Publications Ltd.
- Caers, J. (2011) *Modeling Uncertainty in the Earth Sciences*. 2011 edn. Chichester, West Sussex: Wiley-Blackwell.
- 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., Urquiza, 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.
- Carlsson, G. and Mémoli, F. (2010) 'Multiparameter Hierarchical Clustering Methods', in Locarek-Junge, H. and Weihs, C. (eds.) *Classification as a Tool for Research*. Springer Berlin Heidelberg, pp. 63-70.
- Chang, K.-T. (2012) *Introduction to geographic information systems*. 6th edn. New York: McGraw-Hill.
- Coombes, M. and Raybould, S. (2001) 'Public policy and population distribution: developing appropriate indicators of settlement patterns', *Environment and Planning C: Government and Policy*, 19(2), pp. 223-248.
- Cressie, N. (1990) 'The origins of kriging', *Mathematical Geology*, 22(3), pp. 239-252.
- DCLG (2009) *English Housing Survey, 2009: Housing Stock Data* (SN: 6804). Colchester, Essex: UK Data Archive. [Online]. Available at:

<http://discover.ukdataservice.ac.uk/catalogue/?sn=6804&type=Data%20catalogue>.
DCLG (2010) *English Housing Survey Household report 2008–09* (ISBN 978-1-4098-2600-2). London: DCLG. [Online]. Available at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6695/1750765.pdf (Accessed: 14 July 2013).
DCLG (2011) *English Housing Survey Housing Stock Report 2009* (ISBN: 978-1-4098-3019-1). London: DCLG Publications. [Online]. Available at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6725/1937212.pdf (Accessed: 26 September 2013).
DECC (2008) 'Domestic Energy Consumption, 2008'. London: DECC. Available at:
<http://www.neighbourhood.statistics.gov.uk/dissemination/MetadataDownloadPDF.do?downloadId=26286> (Accessed: 19 June 2013).
DECC (2010) *National Renewable Energy Action Plan for the United Kingdom*. London: DECC. [Online]. Available at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/47871/25-nat-ren-energy-action-plan.pdf.
DECC (2012) *Annex C: Domestic Consumption* (URN 12D/405). London: DECC. [Online]. Available at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/65972/6873-need-report-annex-c.pdf (Accessed: 03 July 2013).
DECC (2013) *Special feature – NEED analysis: Areas and types of properties off the gas grid*. London: DECC. [Online]. Available at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/267375/off_gas_grid.pdf (Accessed: 02 April 2014).
DECC (2014a) 'Data tables: Green Deal, ECO and Insulation Levels, up to December 2013'. London: DECC. Available at: <https://www.gov.uk/government/statistics/green-deal-energy-company-obligation-eco-and-insulation-levels-in-great-britain-quarterly-report-to-december-2013> (Accessed: 04 November 2014).
DECC (2014b) *Sub-national consumption statistics: Methodology and guidance booklet* (URN 12D/473). London: DECC. [Online]. Available at:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/359302/subnational_methodology_and_guidance_booklet.pdf.
DOE (2013) *Getting Started with EnergyPlus: Basic Concepts Manual - Essential Information You Need about Running EnergyPlus*. Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory. [Online]. Available at:
<http://apps1.eere.energy.gov/buildings/energyplus/pdfs/gettingstarted.pdf>.
Dorofeev, S. and Grant, P. (2006) *Statistics for Real-Life Sample Surveys*. Cambridge, UK: Cambridge University Press.
Druckman, A. and Jackson, T. (2008) 'Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model', *Energy Policy*, 36(8), pp. 3177-3192.
Efron, B. and Gong, G. (1983) 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *The American Statistician*, 37(1), pp. 36-48.
Espey, J. and Espey, M. (2004) 'Turnig on the Lights: A Meta-Analysis of Residential Electricity Demand Elasticities', *Journal of Agricultural and Applied Economics*, 36(1), pp. 65-81.
Everitt, B., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*. John Wiley & Sons, Ltd. Available at:
<http://serverlib.moe.gov.ir/documents/10157/42675/Cluster+Analysis.pdf> (Accessed: 23 April 2014).

Ewing, R. and Rong, F. (2008) 'The impact of urban form on U.S. residential energy use', *Housing Policy Debate*, 19(1), pp. 1-30.

Firth, S.K., Lomas, K.J. and Wright, A.J. (2010) 'Targeting household energy-efficiency measures using sensitivity analysis', *Building Research & Information*, 38(1), pp. 25-41.

Granda, P. and Blasczyk, E. (2011) 'XIII. Data Harmonization', *Guidelines for Best Practice in Cross-Cultural Surveys: Comparative Survey Design and Implementation (CSDI) Guidelines Initiative*.

Hamilton, I.G., Steadman, P., Lowe, R., Summerfield, A.J. and Bruhns, H. (2011) *Report on the exploratory analysis of the Homes Energy Efficiency Database and Energy Demand*. London, UK: Buildings and Energy Data Frameworks UCL Energy Institute. [Online]. Available at: <http://www.bartlett.ucl.ac.uk/energy/research/themes/buildings/building-and-energy-data-frameworks> (Accessed: 21 July 2014).

Hiremath, R.B., Shikha, S. and Ravindranath, N.H. (2007) 'Decentralized energy planning; modeling and application—a review', *Renewable and Sustainable Energy Reviews*, 11(5), pp. 729-752.

Hsu, D. (2015) 'Identifying key variables and interactions in statistical models of building energy consumption using regularization', *Energy*, 83(2015), pp. 144-155.

Hughes, M. (2011) *A guide to the Cambridge Housing Model*. London: DECC and CAR. [Online]. Available at: <https://www.gov.uk/government/statistics/cambridge-housing-model-and-user-guide> (Accessed: 02 July 2013).

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.

Jolliffe, I.T. (2002) *Principal Component Analysis*. Springer-Verlag New York. Available at: <http://www.springer.com/us/book/9780387954424>.

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z. and Djurovic-Petrovic, M. (2010) 'A review of bottom-up building stock models for energy consumption in the residential sector', *Building and Environment*, 45(7), pp. 1683-1697.

Kelly, S., Crawford-Brown, D. and Pollitt, M.G. (2012) 'Building performance evaluation and certification in the UK: Is SAP fit for purpose?', *Renewable and Sustainable Energy Reviews*, 16(9), pp. 6861-6878.

Korolija, I. and Zhang, Y. (2013) 'Impact of model simplification on energy and comfort analysis of dwellings', *Proceedings of BS2013: 13th Conference of IBPSA*. Cambéry - France, 26-28 August 2013. Toronto, Canada: IBPSA, pp. 1184-1192. Available at: http://www.ibpsa.org/proceedings/BS2013/p_1502.pdf.

Kusuda, T. (1977) 'Fundamentals of Building Heat Transfer', *Journal of Research of the National Bureau of Standards* 82(2), pp. 97-106.

LGIH (2010) *Data Entry Conventions and Best Practice for the National Land and Property Gazetteer (DEC-NLPG Version 3.1) A Reference Manual*. London: LGIH. [Online]. Available at: <http://www.nlpg.org.uk/nlpg/document.htm?targ=844>.

Li, J. and Heap, A.D. (2008) *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia.

Lin, H.F. (2012) 'A meta-analysis on preschool evaluation researches in Taiwan', *International Journal of Research Studies in Education*, 1(1), pp. 107-114.

Lindberg, R., Korpi, M. and Vinha, J. (2008) 'Factors affecting energy consumption of buildings', *Building Enclosure Science & Technology*. Minneapolis. Available at: https://c.ymcdn.com/sites/www.nibs.org/resource/resmgr/BEST/BEST1_022.pdf.

Ma, J. and Cheng, J.C.P. (2016) 'Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology', *Applied Energy*,

183(Supplement C), pp. 182-192.

McClintock, C.C., Brannon, D. and Maynard-Moody, S. (1979) 'Applying the Logic of Sample Surveys to Qualitative Case Studies: The Case Cluster Method', *Administrative Science Quarterly*, 24(4), pp. 612-629.

National Grid (2012) *Gas Demand Forecasting Methodology*. London: National Grid plc. [Online]. Available at: <http://www.nationalgrid.com/NR/rdonlyres/71CFD0F6-3607-474B-9F37-0952404976FB/52071/GasDemandForecastingMethodologyFeb12.pdf>.

Palmer, J. and Cooper, I. (2012) *United Kingdom housing energy fact file*. [Online].

Available at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/201167/uk_housing_fact_file_2012.pdf.

Pattison, B., Diane Diacon, D. and Vine, J. (2010) *Tenure Trends in the UK Housing System: Will the private rented sector continue to grow?* [Online]. Available at:

http://downloads2.belvoirlettings.com/CentralOffice/BSHF_Tenure_Trend_Research-FINAL_complete.pdf.

Rafferty, A. (2011) *Introduction to Complex Sample Design in UK Government Surveys* [Online]. Available at:

<http://www.esds.ac.uk/government/docs/complexsampledesign.pdf>.

Robinson, J.A. (2014) 'Quantifying polysemy in Cognitive Sociolinguistics', in Glynn, D. and Robinson, J.A. (eds.) *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*. John Benjamins Publishing Company.

Rowntree, M. (2010) *Socio-Economic Profiling for Newcastle upon Tyne: Newcastle City Council - Analysis of MOSAIC Groups, 2009*. Newcastle upon Tyne, UK: Newcastle City Council. [Online]. Available at:

<http://www.newcastle.gov.uk/wwwfileroot/legacy/cxo/instantatlas/Equalities%20Framework/Mosaic%20Equalities.pdf>.

Rylatt, R.M., Gadsden, S.J. and Lomas, K.J. (2003) 'Methods of predicting urban domestic energy demand with reduced datasets: a review and a new GIS-based approach', *Building Services Engineering Research & Technology*, 24(2), pp. 93-102.

Shove, E. and Walker, G. (2014) 'What Is Energy For? Social Practice and Energy Demand', *Theory, Culture & Society*.

Stake, R.E. (1978) 'The Case Study Method in Social Inquiry', *Educational Researcher*, 7(2), pp. 5-8.

Stone, B. and Rodgers, M.O. (2001) 'Urban Form and Thermal Efficiency: How the Design of Cities Influences the Urban Heat Island Effect', *Journal of the American Planning Association*, 67(2), pp. 186-198.

Sui, D.Z. (2004) 'Tobler's First Law of Geography: A Big Idea for a Small World?', *Annals of the Association of American Geographers*, 94(2), pp. 269-277.

Sunikka-Blank, M. and Galvin, R. (2012) 'Introducing the rebound effect: the gap between performance and actual energy consumption', *Building Research & Information*, 40(3), pp. 260-273.

Triantakoustantis, D.P. and Barr, S.L. (2009) 'A Spatial Structural and Statistical Approach to Building Classification of Residential Function for City-Scale Impact Assessment Studies', in Gervasi, O., Taniar, D., Murgante, B., Laganà, A., Mun, Y. and Gavrilova, M. (eds.) *Computational Science and Its Applications – ICCSA 2009*. Springer Berlin Heidelberg, pp. 221-236.

Trigaux, D., Oosterbosch, B., De Troyer, F. and Allacker, K. (2017) 'A design tool to assess the heating energy demand and the associated financial and environmental impact in neighbourhoods', *Energy and Buildings*, 152(Supplement C), pp. 516-523.

UK Data Service (2012) *EHS 2009: Housing Stock Data*. Available at:

<http://ukdataservice.ac.uk/> (Accessed: 27 June 2013).

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.

Vartholomaïos, A. (2017) 'A parametric sensitivity analysis of the influence of urban form on domestic energy consumption for heating and cooling in a Mediterranean city', *Sustainable Cities and Society*, 28(Supplement C), pp. 135-145.

Washington State University (2007) *Capital Planning and Development-Space Management Building area measurements*. [Online]. Available at:

<http://access.ewu.edu/Documents/Construction%20and%20Planning/Campus%20Space%20Data/BuildingAreaMeasurement.pdf>.

Yeo, I.-A., Yoon, S.-H. and Yee, J.-J. (2013) 'Development of an Environment and energy Geographical Information System (E-GIS) construction model to support environmentally friendly urban planning', *Applied Energy*, 104(Supplement C), pp. 723-739.