Challenges of working with a large building energy database. Combining data sets from different scales

Javier Urquizo[a], Carlos Calderón[b] and Philip James[c]

[a] Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador, jurquizo@espol.edu.ec
[b] Newcastle University, Newcastle upon Tyne, United Kingdom, carlos.calderon@newcastle.ac.uk
[c] Newcastle University, Newcastle upon Tyne, United Kingdom, philip.james@newcastle.ac.uk

Abstract:
One of the important issues in geographical analysis is the scale at which the analysis is performed, as different processes may operate at different levels of geographical aggregation in the data set. In this paper, we have used the lowest level of areal aggregation obtainable as the smallest unit of analysis, the single building. Even so, the case study used in this paper demonstrates that including higher levels of geographical aggregation simultaneously in a model of smaller units is essential to draw useful energy consumption conclusions from the data analysis. We use the top-down approach and the bottom-up approach to give a better description of the smallest unit of analysis. The top-down approach uses object representation learned from examples to detect an object in each input building and provide an approximation to its figure-attribute-ground representation from national surveys. The bottom-up approach uses an archetype-based criterion from local surveys to define coherent groups of buildings that are likely to belong together to an urban texture. The combination provides a final urban image that draws on the relative merits of both approaches. The result is as close as possible to the top-down approximation but is also constrained by the bottom-up process to be consistent with significant urban texture discontinuities. Our experiments seem to suggest that individual building extended archetype records derived from an augmentation algorithm are superior to results given by a pure top-down or pure bottom-up approaches.

Keywords:

Urban Energy Models, Urban Texture, Comparison between top-down and bottom-up parsing, Context-free grammars, Finite automation machine, Hamming distance.

1. Introduction

This research inherited the Newcastle Carbon Route Map (NCRM), which is an early incarnation of a building level data set for Newcastle upon Tyne, UK. The initial phase of this research involved substantial data management, cleaning, restructuring and additions to this initial data set. The resultant data set incorporated in a single database table a large number of building related data sets. The Newcastle Carbon Route Map Framework (NCRF) utilises this data set and adds on the energy modelling aspect through linking with the English House Survey (EHS) as input to the Cambridge Housing Model (CHM). This provides the means to produce building level carbon and energy consumption estimates which in turn can be analysed both spatially [1-3] and aspatially (e.g., by building type) [4, 5]. This building level approach through the NCRF provides the potential for energy planners and other bodies to model energy interventions with flexibility in scale and to potentially adapt plans to local area characteristics.

This section outlines the major software components used in NCRF. The software components are used 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 ArcMap™ from ESRI or CHM, or (ii) using executable modules custom made for this study, and called Structured Query Language (SQL) scripts.[1] Fig. 1 and 2 in this section follow a top-down, and cross-functional process flow description diagram. Fig. 1 is the main process flow leading to the NCRF energy consumption estimates – this section follows the natural flow of NCRF i.e., from NCRM data set to the energy estimates − and Fig. 2 is the secondary process flow that creates 104 energy variables from EHS data sets for the record augmentation strategy. Fig. 2 is embedded in Fig. 1 in the process called “EHS energy variables” in violet.

The PostgreSQL™/PostGIS™ software components of the initial NCRM project were selected before this study was initiated. This database was then significantly remodelled, restructured, and new data added. In addition, multiple functional scripts were developed to carry out many of the operations of NCRF. The very first process is the NCRM spatial data formation using the “database joins” and “spatial join” functions.

[1] The scripts −list of commands that are executed to automate processes– are made using PL/SQL. PL/SQL is a loadable procedural language for the PostgreSQL™ database system.

Database joins were used to join data sets with a Unique Property Reference Number (UPRN) shared identifier i.e. Your Homes Newcastle, Registered Social Landlords, and WarmZone data sets, so that their attributes were appended to the NCRM Map primary data set; and spatial joins were used to provide a link (using their common shared location) between the Gazetteer data, the classification data in SCORCHIO and Cities Revealed data sets. At the end of the spatial data formation process a NCRM data set with 122,733 records was in a PostgreSQL^{TM} database, as shown in the top of Fig. 1. This augmented and linked data set is referred to in this paper as NCRM SAP profile. This data set is the basis for all subsequent processing carried out in NCRF.

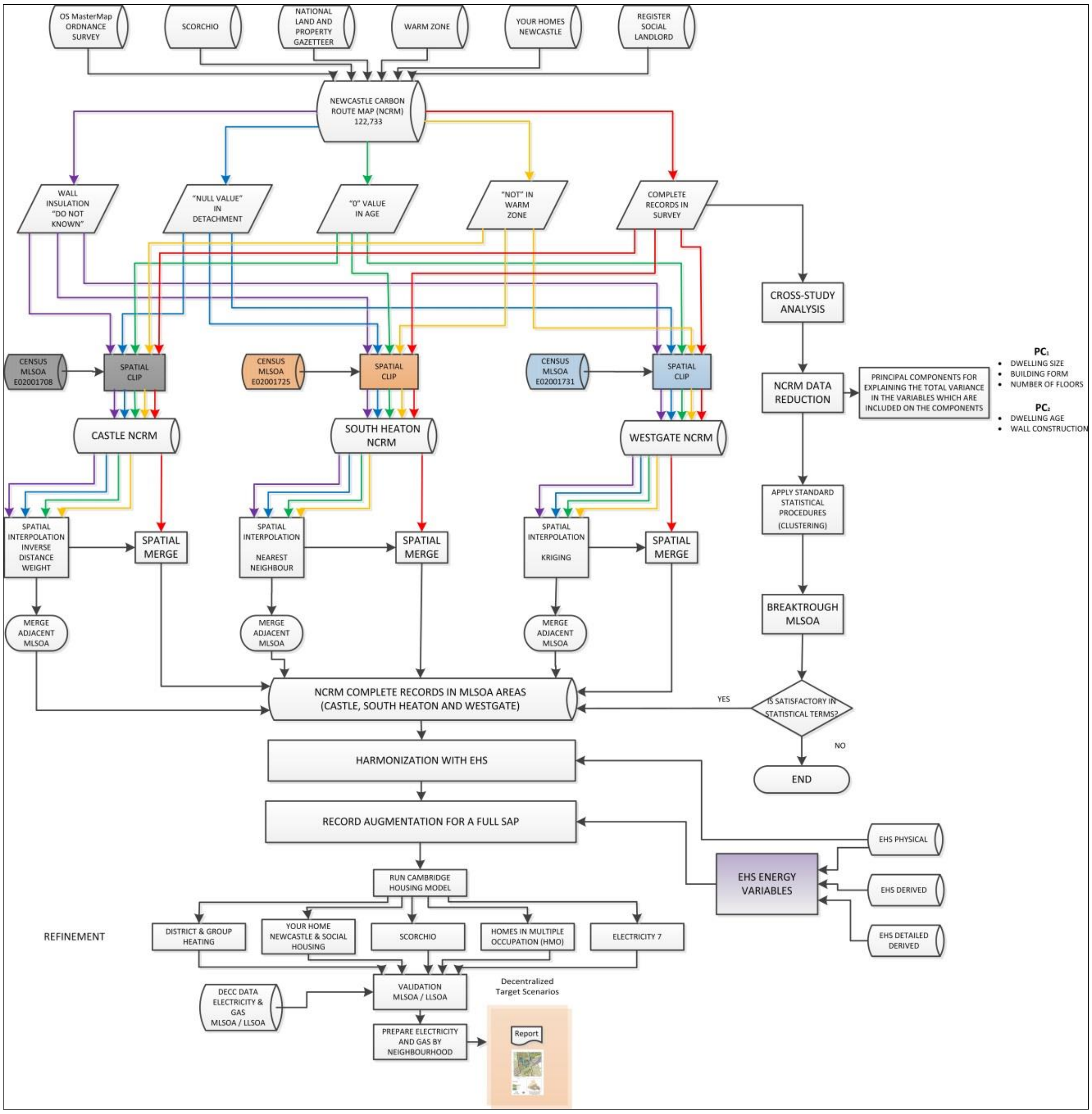

OS MasterMap ORDNANCE SURVEY
SCORCHIO
NATIONAL LAND AND PROPERTY GAZETTEER
WARM ZONE
YOUR HOMES NEWCASTLE
REGISTER SOCIAL LANDLORD
NEWCASTLE CARBON ROUTE MAP (NCRM) 122,733
WALL INSULATION "DO NOT KNOWN"
"NULL VALUE" IN DETACHMENT
"0" VALUE IN AGE
"NOT" IN WARM ZONE
COMPLETE RECORDS IN SURVEY
CROSS-STUDY ANALYSIS
CENSUS MLSOA E02001708
SPATIAL CLIP
CENSUS MLSOA E02001725
SPATIAL CLIP
CENSUS MLSOA E02001731
SPATIAL CLIP
NCRM DATA REDUCTION
PRINCIPAL COMPONENTS FOR EXPLAINING THE TOTAL VARIANCE IN THE VARIABLES WHICH ARE INCLUDED ON THE COMPONENTS
PC1
DWELLING SIZE
BUILDING FORM
NUMBER OF FLOORS
PC2
DWELLING AGE
WALL CONSTRUCTION
CASTLE NCRM
SOUTH HEATON NCRM
WESTGATE NCRM
APPLY STANDARD STATISTICAL PROCEDURES (CLUSTERING)
SPATIAL INTERPOLATION INVERSE DISTANCE WEIGHT
SPATIAL MERGE
SPATIAL INTERPOLATION NEAREST NEIGHBOUR
SPATIAL MERGE
SPATIAL INTERPOLATION KRIGING
SPATIAL MERGE
BREAKTROUGH MLSOA
MERGE ADJACENT MLSOA
MERGE ADJACENT MLSOA
MERGE ADJACENT MLSOA
NCRM COMPLETE RECORDS IN MLSOA AREAS (CASTLE, SOUTH HEATON AND WESTGATE)
YES
IS SATISFACTORY IN STATISTICAL TERMS?
NO
HARMONIZATION WITH EHS
END
RECORD AUGMENTATION FOR A FULL SAP
EHS PHYSICAL
EHS ENERGY VARIABLES
EHS DERIVED
EHS DETAILED DERIVED
RUN CAMBRIDGE HOUSING MODEL
REFINEMENT
DISTRICT & GROUP HEATING
YOUR HOME NEWCASTLE & SOCIAL HOUSING
SCORCHIO
HOMES IN MULTIPLE OCCUPATION (HMO)
ELECTRICITY 7
DECC DATA ELECTRICITY & GAS MLSOA / LLSOA
VALIDATION MLSOA / LLSOA
Decentralized Target Scenarios
PREPARE ELECTRICITY AND GAS BY NEIGHBOURHOOD
Report

Fig. 1. Newcastle Route Map Framework (NCRF)

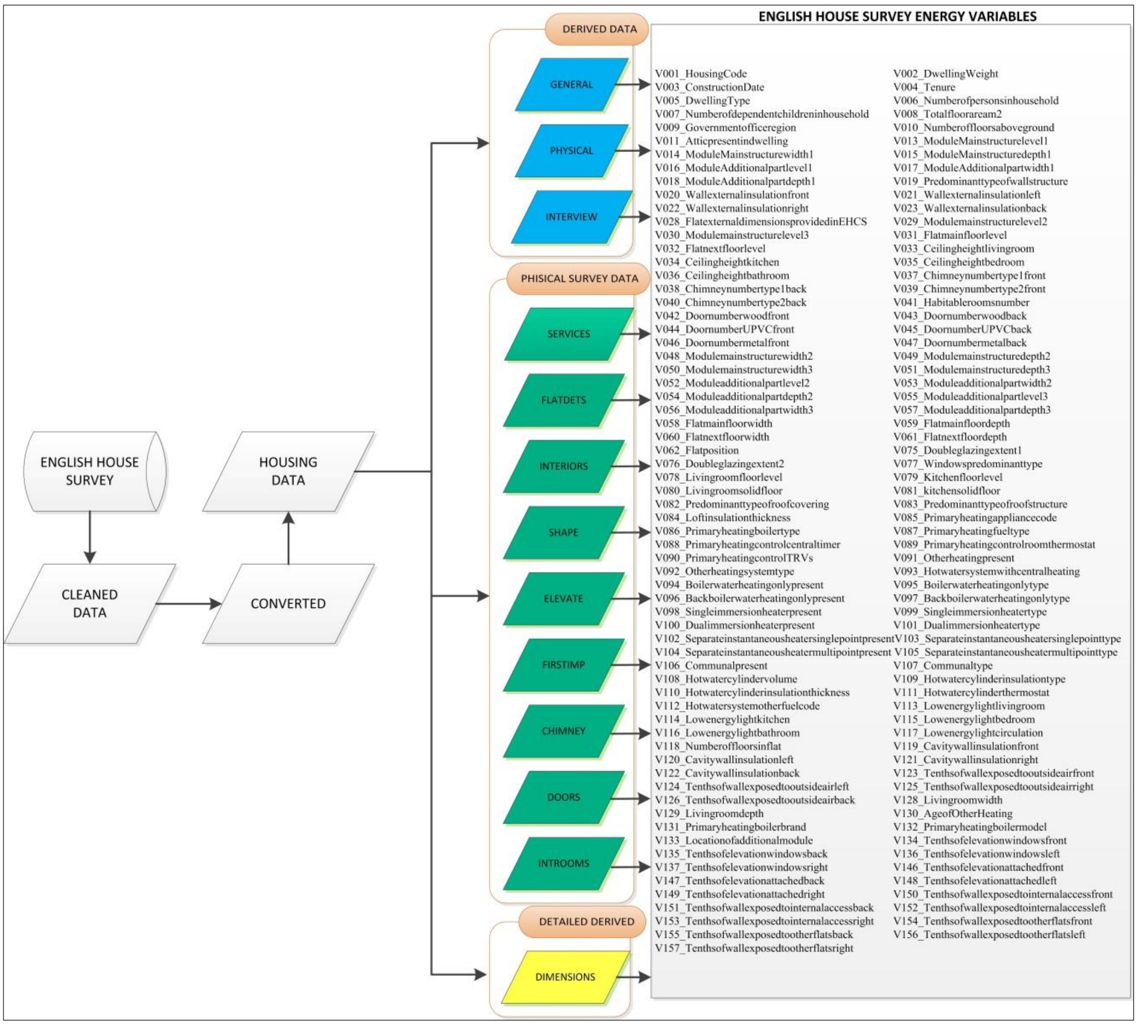
The first activity in the NCRF general process is the cross-study analysis. This activity creates a statistical combination of the NCRM and EHS studies. For the NCRM data set, it involves, first, the separation of the "complete dwelling records" from the "incomplete dwelling records". NCRM incomplete records are records with one or more the following conditions: "do not know" in the wall insulation field, "null value" in detachment field, and '0' value in the dwelling age field. The NCRM data set of the complete records is labelled NCRM cluster profile in Fig. 1. The number of NCRM cluster profile records is 60,977 out of 78,475 dwelling polygon features of the NCRM WarmZone data set; therefore, the incomplete records are 17,498. The separation process is done using (SQL) scripts. The total number of city-records is 122,733, for the difference (122,733 – 78,475 = 44,258 dwellings), or records not in the NCRM WarmZone data set, all fields are unknowns.

Cross-study analysis needs first to harmonize the data sets between NCRM and EHS, second is to query records that contain six categorical variables – dwelling age, wall construction, building form, dwelling size, heating, and number of floors − where dwelling age is based in bands. The third activity is the NCRM data reduction through the principal component analysis (PC) and factor analysis (FA) (see to the right of Fig. 1). The PC and FA are modules of the Statistical Package for Social Science (SPSS^{TM}) software. The PCs are shown in the right side of Fig. 1. The last activities in the right of Fig. 1 is the creation of clusters using a twostep approach – using the hierarchical and k-means clustering techniques – also using the SPSS™ statistical package and the cluster spread in the MLSOAs areas – using SQL scripts.

At this point, the three Middle Layer Super Output Areas (MLSOA) of the case study: Castle, South Heaton and Westgate were selected. PostgreSQL™ records were published to PostGIS™ projects using the geometry field. The spatial extents of every MLSOA were created using the 'clip spatial'[2] component in PostGIS™. The PostGIS™ spatial representation of the three MLSOA case studies was exported to ARCMap™ projects for further analysis. The record generation technique used was spatial interpolation. Spatial Interpolation components were used from the ArcMap™ ArcToolbox. The line colours in the Fig. 1 represent the source of the missing field. For dwellings outside the interpolation area an additional activity is made by merging neighbouring area and then reapplying the same spatial interpolation components. Similar components were applied for the dwelling records outside the NCRM WarmZone data set. ArcMap™ was used as the principal development tool for this stage of the research.

The purpose of the record generation process is to have complete coverage of dwelling records to compare with government statistics in MLSOA and Lower Layer Super Output Areas (LSOAs). After the record generation process, the three spatial data sets of the case study areas were imported back to PostgreSQL™ using SQL scripts, for the record augmentation process. MLSOAs on average have a population of 7,200 whereas LLSOAs have on average a population of 1,500. The record augmentation process is described in detail in Section 2 of this paper.

In a parallel data processing operation, and before data from the English Housing Survey (EHS) can be used in an SAP-based model like the Cambridge Housing Model (CHM), EHS data sets need to be cleaned and converted to align it with the inputs needed for the Standard Assessment Procedure (SAP). SAP is the methodology used by the UK Government to assess and compare the energy and environmental performance of dwellings. A unique record is made from the different EHS data sets: Derived, Physical Survey and Detailed Derived data sets (see Fig. 2).



ENGLISH HOUSE SURVEY ENERGY VARIABLES

V001_HousingCode V002_DwellingWeight
V003_ConstructionDate V004_Tenure
V005_DwellingType V006_Numberofpersonsinhousehold
V007_Numberofdependentchildreninhousehold V008_Totalfloorarеam2
V009_Governmentofficeregion V010_Numberoffloorsaboveground
V011_Atticpresentindwelling V013_ModuleMainstructurelevel1
V014_ModuleMainstructurewidth1 V015_ModuleMainstructuredepth1
V016_ModuleAdditionalpartlevel1 V017_ModuleAdditionalpartwidth1
V018_ModuleAdditionalpartdepth1 V019_Predominanttypeofwallstructure
V020_Wallexternalinsulationfront V021_Wallexternalinsulationleft
V022_Wallexternalinsulationright V023_Wallexternalinsulationback
V028_FlatexternaldimensionsprovidedinEHCS V029_Modulemainstructurelevel2
V030_Modulemainstructurelevel3 V031_Flatmainfloorlevel
V032_Flatnextfloorlevel V033_Ceilingheightlivingroom
V034_Ceilingheightkitchen V035_Ceilingheightbedroom
V036_Ceilingheightbathroom V037_Chimneynumbertype1front
V038_Chimneynumbertype1back V039_Chimneynumbertype2front
V040_Chimneynumbertype2back V041_Habitableroomsnumber
V042_Doornumberwoodfront V043_Doornumberwoodback
V044_DoornumberUPVCfront V045_DoornumberUPVCback
V046_Doornumbermetalfront V047_Doornumbermetalback
V048_Modulemainstructurewidth2 V049_Modulemainstructuredepth2
V050_Modulemainstructurewidth3 V051_Modulemainstructuredepth3
V052_Moduleadditionalpartlevel2 V053_Moduleadditionalpartwidth2
V054_Moduleadditionalpartdepth2 V055_Moduleadditionalpartlevel3
V056_Moduleadditionalpartwidth3 V057_Moduleadditionalpartdepth3
V058_Flatmainfloorwidth V059_Flatmainfloordepth
V060_Flatnextfloorwidth V061_Flatnextfloordepth
V062_Flatposition V075_Doubleglazingextent1
V076_Doubleglazingextent2 V077_Windowspredominanttype
V078_Livingroomfloorlevel V079_Kitchenfloorlevel
V080_Livingroomsolidfloor V081_kitchensolidfloor
V082_Predominanttypeofroofcovering V083_Predominanttypeofroofstructure
V084_Loftinsulationthickness V085_Primaryheatingappliancecode
V086_Primaryheatingboilertype V087_Primaryheatingfueltype
V088_Primaryheatingcontrolcentraltimer V089_Primaryheatingcontrolroomthermostat
V090_PrimaryheatingcontrolTRVs V091_Otherheatingpresent
V092_Otherheatingsystemtype V093_Hotwatersystemwithcentralheating
V094_Boilerwaterheatingonlypresent V095_Boilerwaterheatingonlytype
V096_Backboilerwaterheatingonlypresent V097_Backboilerwaterheatingonlytype
V098_Singleimmersionheaterpresent V099_Singleimmersionheatertype
V100_Dualimmersionheaterpresent V101_Dualimmersionheatertype
V102_SeparateinstantaneousheaterssinglepointpresentV103_Separateinstantaneousheaterssinglepointtype
V104_Separateinstantaneousheatermultipointpresent V105_Separateinstantaneousheatermultipointtype
V106_Communalpresent V107_Communaltype
V108_Hotwatercylindervolume V109_Hotwatercylinderinsulationtype
V110_Hotwatercylinderinsulationthickness V111_Hotwatercylinderthermostat
V112_Hotwatersystemotherfuelcode V113_Lowenergylightlivingroom
V114_Lowenergylightkitchen V115_Lowenergylightbedroom
V116_Lowenergylightbathroom V117_Lowenergylightcirculation
V118_Numberoffloorsinflat V119_Cavitywallinsulationfront
V120_Cavitywallinsulationleft V121_Cavitywallinsulationright
V122_Cavitywallinsulationback V123_Tenthsofwallexposedtooutsideairfront
V124_Tenthsofwallexposedtooutsideairleft V125_Tenthsofwallexposedtooutsideairright
V126_Tenthsofwallexposedtooutsideairback V128_Livingroomwidth
V129_Livingroomdepth V130_AgeofOtherHeating
V131_Primaryheatingboilerbrand V132_Primaryheatingboilermodel
V133_Locationofadditionalmodule V134_Tenthsofelevationwindowsfront
V135_Tenthsofelevationwindowsback V136_Tenthsofelevationwindowsleft
V137_Tenthsofelevationwindowsright V146_Tenthsofelevationattachedfront
V147_Tenthsofelevationattachedback V148_Tenthsofelevationattachedleft
V149_Tenthsofelevationattachedright V150_Tenthsofwallexposedtointernalaccessfront
V151_Tenthsofwallexposedtointernalaccessback V152_Tenthsofwallexposedtointernalaccessleft
V153_Tenthsofwallexposedtointernalaccessright V154_Tenthsofwallexposedtootherflatsfront
V155_Tenthsofwallexposedtootherflatsback V156_Tenthsofwallexposedtootherflatsleft
V157_Tenthsofwallexposedtootherflatsright

Fig. 2. Formation of the EHS energy variables

[2] A command that extracts features from one feature class (NCRM) that reside entirely within a boundary defined by features in another feature class (the UK census MLSOA).

The EHS data sets originally in the excel format were converted to an intermediate format, and then imported to the PostgreSQL™ format using SQL scripts. At the end of these processes 16,150 EHS energy records with 115 fields were available in a PostgreSQL™ database, see violet box in Fig. 1.

However, a further process is required to actually compute the residential area of complex mixed-use buildings. For this the NCRM SCORCHIO data was used augmented by painstaking analysis of online mapping. The result is a reasonably accurate floor area estimate for domestic buildings.

NCRM spatial records were then harmonized with EHS attributes through field attribute mapping and domain mapping of attribute values. This is done through a series of SQL scripts.

2. The record augmentation process

The MCRM record augmentation process to a full SAP is made through Multiple imputations (MI). MI [6, 7] is a practical method for valid inferences for unknown values i.e. filling missing data with plausible values. [8] propose a fundamental shift in the imputation techniques in the way that observed values provide indirect evidence about the likely values of the unobserved ones and [9] use a procedure called data augmentation closely related to their work. This research uses multiple imputations as a data augmentation procedure in which the NCRM records are augmented by EHS records. Augmentation strategy predicts values for urban areas where data is missing (the technique applied is nearest neighbour multiple imputation). Missing responses in the local NCRM survey might be estimated (or imputed) as being the most common answer given by other similar respondents in the EHS survey.

We argue that Principal Components Analysis (PCA) (reducing the data to a smaller number of components) [10], and Factor Analysis (FA) could be used to understand which constructs underlie the NCRM archetype parameters [10]. Then PCA and FA may 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 BREDEM-8 model is used in CHM as a part of the NCRF. 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 factors (see Table 1), tenure is also used. It has been argued [11] that important differences in stock condition and energy efficiency can be identified between the three main tenures in the UK: owner occupation, private renting and social renting. The full set of common fusion variables (CFV)[3] [12] is shown in Table1. 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 [13]. This process is also referred to in the literature as data merging or data consolidation.

Table1. Common fusion variables

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
finmhfue	Main heating fuel
Finchtyp	Primary heating system - type of system
finmhboi	Boiler group
tenure8x	Tenure

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 using a bottom-up approach, 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 in a top-down approach. The approximate matching problem is based on the edit-distance between two strings. There is abundant literature in the field of computer science on how this can be achieved. Section 2 highlights the concepts and the scientific advancements in the field of linguistics and the many opportunities that potential uses for shape grammars (as a means of abstraction) for architects and engineers interested in the field of sustainable energy systems. Section 2 introduces the methodology for a given context-free grammar (CFG) i.e., a NCRM record and a finite-state automaton (FA), i.e., the EHS

[3] The technique was used mainly as a way of imputing missing individual data items.

record and the edit-distance problem solves the problem of computing the most similar pair of strings between a local and national data set.

2.1 Record augmentation using an approximate match between NCRM an EHS

The basic parameterizable rules appears in architecture and eventually led to the shape grammars [14, 15]. Later, the ideas of shape grammars were brought into computer graphics [16, 17] The initial emphasis was on creating grammars that could generate a set of buildings for computer graphics, although later work extended the ideas to visual editing in an attempt to allow the design of new buildings [18]. Since shape grammars are built on top of context-free grammars (CFG), we briefly review CFGs here. A context-free grammar consists of a finite set of symbols and a set of productions P mapping the symbols to strings of symbols. Often the symbols are separated into variables.

Context-free grammars arise in linguistics where they are used to describe the structure of words and sentences in a natural (human) language, and they were discussed by Noam Chomsky for this purpose. A context-free grammar provides a mathematics for describing the methods by which phrases in natural languages are built from smaller blocks, capturing the "block structure" of sentences in a natural way. Its simplicity makes the formalism amenable to rigorous mathematical study. The "block structure" aspect that context-free grammars capture is so fundamental to grammar that the terms syntax and grammar are often identified with context-free grammar rules, especially in computer science. Likewise, building archetypes are theoretical buildings created by a composite of several characteristics found within a category of buildings with similar attributes supporting a laborious manual architectural and engineering design of rules. Learning is based on a large local building data set covering detailed characteristics of the house (e.g., dwelling size, geometry, and characteristics of the building envelope), the energy (heating) system, the household characteristics (e.g., income, size and composition), the climate (e.g., external temperature, monthly average wind speed, and monthly average horizontal solar radiation), the landscape (e.g., urban form, plot ratio) and their interrelations. In particular, a cluster has been derived from a local data set and employed for structure learning. Therefore, an archetype is a virtual representation of a number of buildings that share similar characteristics in a building stock. Context-free grammars are simple enough to allow the construction of efficient parsing algorithms that, for a given string, determine whether and how it can be generated from the grammar. By contrast, in computer science, as the use of recursively defined concepts increased, context-free grammars were used more and more.

A gap cost model for the edit-(linear)distance is crucial for finding a proper alignment between two sequences. In this research we design an efficient algorithm for the edit-distance between a CFG and an FA under a gap cost model. A FA is a simple idealized machine used to recognize patterns within input taken from some character set (or building archetype). The FA searches an input depending on whether the pattern defined by the FA (i.e., in the EHS) occurs in the input.

The general theoretical literature on combining data sets from different scales in the United Kingdom is inconclusive on several vital questions within the energy discourse. We examine the edit-distance between two data sets at different scales based in similar problems experienced in context-fee grammars in computer science. There is evidence from the existing literature on how to compute the edit-distance. Examples are: [19] provided a quadratic time algorithm for computing the edit-distance between two regular languages. [20] considered the relative edit-distance between languages and the reflexivity of binary relations. [21] considered the minimum edit-distance among all pairs of distinct strings of the language.

2.2 Record augmentation conceptual foundation

Let us say that a CRM record is a 10-tuple w (see Table 1) in a set L(G) for some context-free grammar (CFG) G has been parsed when we know one (or perhaps all) of its derivation trees. This tree may be physically constructed in the computer memory. One can deduce the parse tree by watching the steps taken by a syntax analyser. A syntax analyser or parser takes the input from a lexical analyser in the form of token streams. The parser analyses the token stream against the production rules to detect any errors in the code. The output of this phase is a parse tree.

Most parsers simulate a Pushdown Automata PDA which is recognizing the input either top-down or bottom-up. The ability of a PDA to parse top-down is associated with the ability to map input strings to their leftmost derivations whereas bottom-up parsing is associated with mapping input strings to their rightmost derivations. The parsing problem is therefore mapping strings to leftmost of rightmost derivations. While there are many other parsing strategies, these two definitions are the significant ones [22]. It was decided that the best parsing model to adopt for this investigation is the bottom-up approach, the reason is in the following paragraph.

In the bottom-up detailed parts are at the bottom of the upside-down tree, and larger structures composed from them are in successively higher layers, until at the top or "root" of the tree a single unit describes the entire input stream. A bottom-up parse discovers and processes that tree starting from the bottom left end, and incrementally works its way upwards and rightwards. A parser may act on the structure hierarchy's levels without ever creating an actual data tree; the tree is then merely implicit in the parser's actions.

However, there is an issue underpinning this method that must be considered. The unit distance of measure in categorical attributes is diverse. The notion of similarity or distance for categorical data is not as straightforward as for continuous data. The key characteristic of categorical data is that the different values that a categorical attribute takes are not inherently ordered. Thus, it is not possible to directly compare two different categorical values [23]. Additionally, the magnitude (or distance) between the full categorical score values (the range) is different and has no origin. For example, the categorical range in the heating variable is 16 i.e., the different combinations of fuel type and heating systems, while the categorical range of wall construction variable is six i.e., the different combinations of wall type and wall insulation, and a categorical value of 15 in the variable heating system is not necessarily fifteen more efficient than a categorical value of one, so does the same apply for wall construction. Last, categorical scores have no origin; a score of zero in wall construction does not necessarily imply an absence of the wall. However, for numerical attributes, distance measures are a natural concept. In NCRM a unit distance could mean using a different type of boiler or being in a range in a dwelling size or dwelling age. The impact on the energy consumption is bigger in a unit distance in dwelling size than any other unit distance. Additionally, the order in which the tuples are arranged may affect the results when an algorithm is executed; if a field is biased, the distance measure used between tuples is not perfect.

2.2 Record augmentation mathematical modelling

Given a tuple ω and a text *T*, the string-matching problem is to find occurrences of ω in *T* and the approximate matching problem is to find occurrences of ω' in *T* such that ω and ω' are similar by a predefined similarity metric. The most common similarity metric is the edit-distance [24]. Many researchers investigated the approximate pattern matching problem with various types of mismatches. Examples are: [25] present an algorithm that will parse any input string to completion finding the fewest possible number of errors, [26] develops an algorithm that is similar to both Knuth's LR(k) algorithm and the familiar top-down algorithm, [27] develop a least-errors recognizer using the recognizer of Earley, along with elements of Bellman's dynamic programming, [28] consider efficient methods for computing a difference metric between two sequences of symbols, where the cost of an operation to insert or delete a block of symbols is a concave function of the block's length, [29] generalizes the Cocke-Younger-Kasami algorithm for determining membership in a context-free language, [30] argued that it could be desirable to use concave weighting functions, [31] introduce an attribute matching system and a separate control grammar, which offer the flexibility required to model buildings using a large variety of different styles and design ideas, and [15] established the formal machinery for the algorithmic definition of languages of two- and three-dimensional spatial designs.

In summary, the edit-distance *d(x,y)* between two strings *x* (NCRF) and y (EHS) is the smallest number of operations that transform *x* to *y* i.e. the minimal cost of an alignment w between x and y. Researchers considered different atomic edit operations in different applications in the literature [27]. Here we consider a basic insertion operation. In practical terms, the edit-distance of two strings x and y is:

$$d(x,y) = \min \{c(w)/h(w) = (x,y)\} \qquad (1)$$

We say that *w* is optimal if $d(x,y) = c(w)$

We simply denote the cost $c(\sigma)$ of an individual edit operation $\sigma = (a \rightarrow b)$ by $c(a,b)$ instead of $c = ((a \rightarrow b))$, where $a, b \ \varepsilon\ \Sigma\ U \{\lambda\}$. We assume that the cost function c satisfies two conditions: $c(a,b) = c(b,a)$ and c(a,a) = 0, for a, $b\ \varepsilon\ \Sigma\ U \{\lambda\}$.

3 Main algorithms

The main idea of our algorithm is to compute the edit-distance between two strings [32]. We compute distances for all variables of the CFG and all pairs of states of the FA. Note that this approach is similar to the procedure for constructing a CFG that generates the intersection of a CFG and an FA [18].

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. One of the first activities is the field recoding of NCRM as shown in Table 2, for tenure as an example.

Table 2. NCRF field recode.

Recode	NCRF (tenure)	Sample percentage	EHS (tenure)	Sample percentage
4	Housing Association	4	HA - occupied	15.6
3	Loal Authority	28.9	Local Authority - occupied	13.9
1	Owner Occupied	56.1	Owner Occupied - occupied	50.8
2	Private rented	11	Private rented - occupied	15.8

From Table 2, NCRM Local Authority is high, and Housing Association is low. Therefore, local area characteristics are important to the understanding of the energy consumption estimates in sub-city areas; energy efficiency measures are area specific, and the application must be area-based, and these areas may not align with UK aggregation areas.

Then Algorithm 1 follows to populate the data sets.

Algorithm 1-- Main function for hamming distance - Castle full data set to populate the crm_ehs_pair i.e., its dictionary (MLSOA)

```
create or replace function haming4()
returns setof castle_energy as
declare
-- Declare variables for CRM 'per type analysis'.
        crm_groupid     text;
        crm_toid        text;
        crm_nat_uprn    text;
        crm_floorarea   int;
        .
        .
        .
-- Declare counters
        crmr_count      int;                    -- loop in crm
        ehs_count       int;                    -- loop in ehs
-- Declare records and cursors
        rec_crmr        record;                                         -- record in crm
        cur_crm         Cursor for select * from castle_en;             -- cursor in crm
for variable rec_ncrmf in cursor cur_ncrf
  Loop -- Get one per-property type record - one column at a time.
        select rec_ncrmf.groupid        into crm_groupid        from sheaton_en;
        select rec_ncrmf.toid           into crm_toid           from sheaton_en;
        select rec_ncrmf.nat_uprn       into crm_nat_uprn       from sheaton_en;
        select rec_ncrmf.floorarea      into crm_floorarea      from sheaton_en;
        .
        .
        .
        if (crmr_count>0 and crm_groupid        is not null
                         and crm_Toid           is not null
                         and crm_nat_uprn       is not null
                         and crm_floorarea      <> 0
                         .
                         .
                         .                      ) then
                perform crm_ehs_plot_f1s (crm_Toid, crm_nat_uprn, crm_floorarea, . . .);
        end if;
        -- Return  next rec_crmr;
  End loop;
end;
```

Algorithm 2 is called inside Algorithm 1 to do the d-distance (hamming distance) computation.

Algorithm 2—function crm_ehs_plo_f1s

```
Create or replace function crm_ehs_plot_f1c(crm_toid text, crm_nat_uprn text, crm_floorarea integer…)
  returns setof ehsdataview as
declare
-- Declare variables for English Housing Survey
        CHM_Input       text;   -- English House Survey Code aacode = Cambridge Housing Model aacode
        CHM_Inputb      text;
        CHM_Inputa      text;
```

```
            ehs_floorarea    int;
            .
            .
            .
-- Declare individual distance variables and weight
            d_floorarea_w    int;
            dist_floorarea   int;
            .
            .
            .
-- Declare variables for hamming distance
            v15              integer;
            processed        boolean;         -- flag for applying hamming distance - TRUE if different.
            row_a            boolean;         -- flag for crm record a with non-zero hamming distance
            hamsidta         int;             -- hamming distance of record a
            row_b            boolean;         -- flag for crm record b (next) with non-zero hamming
            distance
            hamsidtb         int;             -- hamming distance of record b (next)
            hamsidt          int;             -- minimum hamming distance = min (hamsidtb - hamsidta)
            crmdist          text;            -- crm record selected for minimum hamming distance
            ehsdist          text;            -- ehs record selected for minimum hamming distance
            MSOA_range       int;             -- range of the EHS to check around MLSOA
            rec_ehsr         record;          -- record for the EHS cursor
            cur_ehs          cursor for select * from EHS2009_Energy_CRM; -- cursor in ehs - 158 variables
begin
-- Initialize flags for record id
 row_a            := 'TRUE';
 row_b            := 'TRUE';
 -- Initialize flag for equal crm and ehr record
 processed        := TRUE;
 -- Initialize individual hamming distance and weight
 d_floorarea_w    := 1;
 dist_floorarea   := 0;
.
.
.
V15               := 0;
            for rec_ehsr in cur_ehs
            loop
               select rec_ehsr.V001_Housingcode into CHM_Input from EHS2009_Energy_CRM;
               select round(rec_ehsr.V008_Totalflooraream2::numeric, 0) into ehs_floorarea from
                       EHS2009_Energy_CRM;
                .
                .
                .
                V15        := V15 + 1;
               MSOA_range := ehs_dwtype7x + … + ehs_Tenure8x;
                hamsidt   := case when (crm_floorarea) <> (ehs_floorarea) then round
                ((d_floorarea_w*abs(crm_floorarea - ehs_floorarea))::numeric, 0) else 0 end +…     + case
                when (crm_tenure8x)  <> (ehs_tenure8x)  then abs(crm_tenure8x  - ehs_tenure8x)
                else 0 end ;
                      if (hamsidt = 0 OR MSOA_range > 47 OR MSOA_range < 11)     then
                                   -- Do nothing if crm and ehs records are equal or MSOA is out of range
                            processed   := 'FALSE';
                     end if;
                      if (processed AND row_b) then
                            if (row_a and V15 < 3) then row_a = 'FALSE'; end if;
                            if (row_b end V15 < 3) then row_b = 'FALSE'; end if;
                            -- slot a
                            hamsidta          := hamsidt;
                            CHM_Inputa        := CHM_Input;
                        end if;
                        if (PROCESSED and row_a and V15 < 3 ) then
```

```
                    if (row_b = 'FALSE') then row_b = 'TRUE'; end if;
                     -- slot b
                            hamsidtb        :=      hamsidt;
                            CHM_Inputb      :=      CHM_Input;
                end if;
                if (processed and row_a = 'FALSE' and V15 < 3) then row_a = 'TRUE'; end if;
                if (processed and hamsidta <> 0 and hamsidtb <> 0) then
                    hamsidt := case when hamsidtb <= hamsidta then hamsidtb else hamsidta end;
                    CHM_Input := case when hamsidtb <= hamsidta then CHM_Inputb else
                    CHM_Inputa end;
                end if;
                processed    := TRUE;               -- initialize flag for next record iteration
                hamsidtb     := hamsidt;
                CHM_Inputb  := CHM_Input ;
                Return next rec_ehsr;
        end loop;
              perform crm_ehs_insertc (crm_nat_uprn, CHM_Inputb, hamsidtb) ;
end;
```

Algorithm 3 – is called from Algorithm 2 creating a unified record for the spatial record

```
create or replace function crm_ehs_insertc(v10 text, v11 text, v12 integer)
  returns integer as
declare
   result integer;
begin
   insert into crm_ehs_pairc (fieldnat_uprn, fieldchm_input, fieldham_dist) values (v10, v11, v12) returning
   v12 into result;
   return result;
end;
  language plpgsql;
alter function crm_ehs_insertc(text, text, integer) owner to postgres;
```

The final output of the NCRF is a spatially enabled data set of residential dwellings with a per-dwelling energy estimate and a large number of energy-related variables. PostGIS™ 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 research ArcMap™ was used to carry out further analysis of the NCRF results.

4. Results

To reach their full efficiency during operation, we developed a comprehensive algorithm that orchestrates all energy parameters in the building. Such a solution also provides an integration point for decentralized energy production systems and smart grid applications. To operate successfully, the system must be aware of a multitude of different energy parameters in order to make energy efficient decisions on behalf of the user.

The augmentation process imputes the nearest neighbour EHS record following an algorithm implemented using two cursors iterating over the tables to search and compare, see Algorithms 1 and 2. This was written as a series of functions in the procedural language for the PostgreSQL database system (PL/SQL). The execution time of the augmentation process is shown in Table 3.

Table 3. Execution time of the record augmentation process

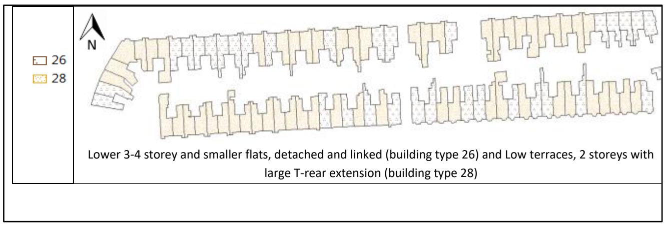
MLSOA	Execution time (ms)	Execution Time (hours)
Castle	24,455,410	6.79
South Heaton	26,667,643	7.41
Westgate	24,598,490	6.83

The final process involves better estimation of the spatial interpolation results, using data sets from the city council regarding districts and group heating, E7, Houses in Multiple Occupations and NCRM SCORCHIO in a model refinement process.

Finally, to compute energy estimates, the NCRM SAP profile data is linked to the CHM. The CHM is written in Microsoft Excel. The CHM was run for all the EHS data (16,150 records). The results of which were imported from CHM into PostgreSQL™. As the full NCRM SAP profile already contains the unique identifier of the best match in the EHS data, a simple lookup is required to pull the matching SAP data from the imported CHM data set.

The resulting values of annual electricity consumption, annual E7 consumption and annual gas consumption were appended to the NCRM spatial record in ArcMap^{TM} using SQL scripts, and become the Newcastle Carbon Route Map Framework (NCRF) energy estimation results.

Figure 3 shows terraces along Simonside Terrace and Rothbury Terrace and Table 4 shows our estimates per Unique Property Reference Number (UPRN) of the total annual Energy (Gas) and Energy (Electricity) consumption in kWh along with the corresponding CO_2 emissions of the Simonside Terrace in tonnes of CO_2. UPRN is a unique numeric identifier for every spatial address in Great Britain. Local authorities have the statutory permission to name and number every street and property in Great Britain and allocate UPRNs.



Lower 3-4 storey and smaller flats, detached and linked (building type 26) and Low terraces, 2 storeys with large T-rear extension (building type 28)

Fig. 3. Linear terraces in South Heaton

Table 4. Per-dwelling energy and carbon estimation of the Simonside Terrace

Address	Area (m^2)	Energy Gas (kWh)	Energy Electricity (kWh)	CO_2 Gas (Tons)	CO_2 Electricity (Tons)
68 Simonside Terrace	86.07	19,050	3,147	3.772	1.627
62 Simonside Terrace	82.79	13,223	2,392	2.618	1.237
64 Simonside Terrace	84.49	19,050	3,174	3.772	1.627
74 Simonside Terrace	98.92	22,287	4,230	4.413	2.187
80 Simonside Terrace	86.59	19,050	3,147	3.772	1.627
84 Simonside Terrace	81.56	17,309	3,373	3.427	1.744
39 Simonside Terrace	93.31	18,190	3,492	3.602	1.805
41 Simonside Terrace	94.23	22,287	4,230	4.413	2.187
43 Simonside Terrace	72.90	11,599	3,115	2.297	1.61
46 Simonside Terrace	89.26	19,050	3,147	3.772	1.627
48 Simonside Terrace	105.24	32,179	2,676	6.371	1.384
50 Simonside Terrace	87.61	19,050	3,147	3.772	1.627
60 Simonside Terrace	102.25	22,287	4,230	4.413	2.187
17 Simonside Terrace	98.09	22,287	4,230	4.413	2.187
19 Simonside Terrace	83.85	19,050	3,147	3.772	1.627
23 Simonside Terrace	91.81	18,190	3,492	3.602	1.805
27 Simonside Terrace	82.85	13,223	2,392	2.618	1.237
78 Simonside Terrace	86.95	19,050	3,147	3.772	1.627
37 Simonside Terrace	94.93	22,287	4,230	4.413	2.187
58 Simonside Terrace	97.83	14,205	3,991	2.813	2.063
76 Simonside Terrace	84.19	19,050	3,147	3.772	1.627
66 Simonside Terrace	121.37	23,207	4,515	4.595	2.334
72 Simonside Terrace	82.56	13,223	2,392	2.618	1.237
82 Simonside Terrace	87.78	19,050	3,147	3.772	1.627
45 Simonside Terrace	50.61	14,630	2,452	2.897	1.793
70 Simonside Terrace	86.27	19,050	3,147	3.772	1.627
15 Simonside Terrace	69.94	19,313	3,467	3.824	1.793
21 Simonside Terrace	88.35	19,050	3,147	3.772	1.627
25 Simonside Terrace	85.05	19,050	3,147	3.772	1.627
29 Simonside Terrace	97.79	22,287	4,230	4.413	2.187
31 Simonside Terrace	86.96	19,050	3,147	3.772	1.627
33 Simonside Terrace	92.93	21,926	5,243	4.341	2.711
35 Simonside Terrace	92.04	18,190	3,492	3.602	1.805
78 A Simonside Terrace	86.95	19,050	3,147	3.772	1.627
80 A Simonside Terrace	86.59	19,050	3,147	3.772	1.627
82 A Simonside Terrace	87.78	28,942	2,389	5.731	1.235
84 A Simonside Terrace	81.56	17,309	3,373	3.427	1.744

As a summary, the main software tools are spatial analysis components from the ArcMap^{TM} ESRI and statistical components from SPSS^{TM}. The SQL scripts components were done using scripts in PostgreSQL^{TM}.

5. DISCUSSION AND FUTURE WORKS

In summary, being a model is a simplified representation of a system, with a dual purpose of enable reasoning within an idealized base scenario framework, and to enable predict of what might happen under energy efficiency measures. NCRF presents an extended SAP building energy profile with a compact explicit simplifying assumption that allows acceptably reasonable simulations of the real interactions of the dwelling, the household characteristics and their close environment and an energy model results that is a better and more close representation of the energy consumption after varying some of those key profile variables. This will allow an initial interaction with reality. From qualitative and quantitative observations of the derived demand of carbon and energy, we might gain a progressive understanding of what other factors are in place in neighbourhoods which seems important. The NCRM spatial approach is flexible to integrate these observations.

One possible application for the NCRF energy consumption are to implement the energy consumption demand for Natural gas Combined Heat and Power (CHP) for districts. Natural gas CHP is the most efficient natural gas generating capacity. Natural gas CHP presents a cost-effective potential for supplying a number of small Heat Networks, with a high proportion of electricity sold to local customers rather than exported to the grid, because the significant barrier of accessing the electricity market wholesale price. Two immediate benefits are: first, networks can make the most of a range of heat sources that individual building solutions cannot harness and second, in urban areas there are challenges with electric heating where building density will prevent wide use of individual ground source heat pumps.

Heat Networks are clearly a significant potential growth area in local governments where a Heat Networks Delivery Unit (HNDU) would be established (as has been the case in Newcastle). HNDU funding of heat mapping and energy master-planning allows Local Authorities to explore and prioritise heat network opportunities through a simple technical and economical assessment. However, detailed project development studies require both a detailed energy consumption model (as the NCRF) and a detail analysis on the carbon intensity of generation displaced by generation from the natural gas CHP capacity.

REFERENCES

[1] Urquizo J, Calderón C, James P, editors. A spatial perspective of the domestic energy consumption intensity patterns in sub-city areas. A case study from the United Kingdom. 2016 IEEE Ecuador Technical Chapters Meeting (ETCM); 2016 12-14 Oct. 2016.

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

[3] Urquizo J, Calderón C, James P. Modelling household spatial energy intensity consumption patterns for building envelopes, heating systems and temperature controls in cities. Applied Energy. 2018;226:670-81.

[4] Urquizo J, Calderón C, James P. Using a Local Framework Combining Principal Component Regression and Monte Carlo Simulation for Uncertainty and Sensitivity Analysis of a Domestic Energy Model in Sub-City Areas. Energies. 2017;10(12).

[5] 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.

[6] Rubin D. Inference and missing data. Biometrika. 1976;63(3):581 - 92.

[7] Schafer JL, Olsen MK. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research. 1998;33(4):545-71.

[8] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological). 1977;39(1).

[9] Tanner MA, Wong WH. The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association. 1987;82(398):528-40.

[10] Everitt B, Landau S, Leese M, Stahl D. Cluster Analysis. West Sussex, UK: John Wiley & Sons, Ltd; 2011. Available from: http://serverlib.moe.gov.ir/documents/10157/42675/Cluster+Analysis.pdf.

[11] Pattison B, Diane Diacon D, Vine J. Tenure Trends in the UK Housing System: Will the private rented sector continue to grow? ; 2010.

[12] Dorofeev S, Grant P. Statistics for Real-Life Sample Surveys. Cambridge, UK: Cambridge University Press; 2006.

[13] Bleiholder J, Naumann F. Data fusion. ACM Comput Surv. 2009;41(1):1-41.

[14] Gips J, Stiny G. Production Systems and Grammars: A Uniform Characterization. Environment and Planning B: Planning and Design. 1980;7(4):399-408.

[15] Stiny G. Introduction to Shape and Shape Grammars. Environment and Planning B: Planning and Design. 1980;7(3):343-51.

[16] Müller P, Vereenooghe T, Wonka P, Paap I, Gool LV. Procedural 3D reconstruction of Puuc buildings in Xkipché. Proceedings of the 7th International conference on Virtual Reality, Archaeology and Intelligent Cultural Heritage; Nicosia, Cyprus: Eurographics Association; 2006. p. 139–46.

[17] Müller P, Wonka P, Haegler S, Ulmer A, Gool LV. Procedural modeling of buildings. ACM SIGGRAPH 2006 Papers; Boston, Massachusetts: Association for Computing Machinery; 2006. p. 614–23.

[18] Bar-Hillel Y, Perles M, Shamir E. On formal properties of simple phrase structure grammars. Jerusalem1960.

[19] Mohri M. Edit-distance of weighted automata: General definitions and algorithms. International Journal of Foundations of Computer Science. 2003;14(06):957-82.

[20] Choffrut C, Pighizzini G. Distances between languages and reflexivity of relations. Theoretical Computer Science. 2002;286(1):117-38.

[21] Konstantinidis S. Computing the edit distance of a regular language. Information and Computation. 2007;205(9):1307-16.

[22] Aho A, Ulman J. The Theory of Parsing, Translation, and Compiling (Volume I: Parsing): Prentice Hall; 1972.

[23] Boriah S, Chandola V, Kumar V. Similarity Measures for Categorical Data: A Comparative Evaluation. SIAM Conference on Data Mining; Atlanta: Society for Industrial and Applied Mathematics; 2008. p. 243-54.

[24] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady. 1966;10:707.

[25] Aho A, Peterson TG. A Minimum Distance Error-Correcting Parser for Context-Free Languages. SIAM J Comput. 1972;1:305-12.

[26] Earley J. An efficient context-free parsing algorithm. Commun ACM. 1970;13(2):94–102.

[27] Lyon G. Syntax-directed least-errors analysis for context-free languages: a practical approach. Commun ACM. 1974;17(1):3–14.

[28] Miller W, Myers EW. Sequence comparison with concave weighting functions. Bulletin of Mathematical Biology. 1988;50(2):97-120.

[29] Myers G. Approximately matching context-free languages. Information Processing Letters. 1995;54(2):85-92.

[30] Waterman MS. Efficient sequence alignment algorithms. Journal of theoretical biology. 1984;108(3):333-7.

[31] Wonka P, Wimmer M, Sillion F, Ribarsky W. Instant architecture. ACM Trans Graph. 2003;22(3):669–77.

[32] Sippu S, Soisalon-Soininen E. Parsing Theory: Volume I: Languages and Parsing. Berlin-Heidelberg-New York: Springer-Verlag; 1988.