THE IDENTIFICATION AND ANALYSIS OF REGIONAL BUILDING STOCK CHARACTERISTICS USING MAP BASED DATA

D.K. Alexander, S. Lannon, O. Linovski
Welsh School of Architecture
Cardiff University
Cardiff, Wales, UK

ABSTRACT

Building energy and Carbon emission calculation methods for regions are of limited use if appropriate input data cannot be economically generated. To enable a wider uptake of regional modelling methods an automated analysis system is required to replace or assist time-consuming and expensive manual surveys of building stock. Building age is an important parameter in estimating energy use and Carbon emissions. In this paper a number of methods to extract information about the built environment from digital maps and use that information to infer building age have been tested against a database of a known large urban region. The methods include different types of shape recognition of plan form and of identification of contextual geography; e.g. distance from entrance to the nearest road.

Tested against samples containing several thousands of domestic buildings from a known region, it was found that the different methods were able to cluster buildings into different form "styles", and that those styles had some correlation to built age. Victorian (pre-1919) age housing was detected with the greatest accuracy, with over 90% in the sample tested correctly identified. This is useful as those older buildings are often the least energy efficient. Success in identification of other eras was less pronounced; although the results are promising, further development of the methods are required.

INTRODUCTION

Boardman et al (2005) estimated that the domestic building stock could deliver a 60% reduction of the UK's carbon emissions by 2050. However, as the proportion of new building stock is small, simply strengthening the building regulations can have only a correspondingly small effect. It is therefore essential to understand the energy performance and carbon emissions of the existing building stock and to be able to predict and test, on a regional level, the potential for improvement.

The Energy and Environment Prediction (EEP) model (Jones, Williams and Lannon, 2000) is a computer based modelling framework that quantifies energy use for regions as an aid to planning actions and incentives aimed at reducing carbon dioxide emissions. The EEP model has proven valuable to its users, for instance in allowing a local authority to target energy efficiency measure to those areas with the greatest need.

However, its use is appropriate only when a region can be described at the level of detail required by the model. Regional energy models such as EEP require an accurate description of the built characteristic of the area in question; as well as numbers and floor areas (both relatively simple to derive) this also includes the profile of housing stock type, age, and fabric characteristics, all required to estimate fuel requirement.

The development of the EEP method showed that an appropriate level of energy and carbon emissions could be achieved by clustering buildings into common types; e.g. "Pre-1919 detached 3-story single family dwelling". Each building in the region to be modelled is then placed in an appropriate cluster in order to estimate its energy use. Building age was determined to be a key concept in forming useful energy performance related clusters; in EEP, building age is defined as the era when the property was built, using date bands commonly used in the UK to describe housing; pre-1919, 1919-1945, 1946-1965, 1966-1980, and post-1980. From the building age characteristic other parameters, such as levels of insulation, uptake of double glazing, or type of heating system, can be inferred.

In application to a real region, this clustering information is acquired at significant expense. Information on built age is, for most local authorities in the UK, not readily available. The data may exist but only in many disparate forms and sources; planning offices, archives, records offices. A manual collation and analysis of those on a large scale (for instance the region considered in this paper comprises 450 square km) would be prohibitively costly. The procedure for the EEP method requires a manual survey to identify the building stock parameters required. This survey is primarily a 'walk by' method, involving the visual identification of each building in the area, coupled with a postal survey covering details such as prevalence of retrofit measures, heating system types and so forth. When this method was applied to the county of Neath Port Talbot in the United Kingdom, where 55,000 dwellings were to be modelled (nearly 100% of the stock population), the survey required an investment of 18 man-months. The author's experiences of this and other surveys have shown that residents are poor judges of their building's age, unless it is very new. We cannot rely on their data alone, and so trained assessors were required.

When well formed, a regional model such as EEP can predict energy use trends with acceptable accuracy, even though the underlying energy model is relatively simple (in the case of EEP, one based on the BREDEM method). Figure 1 shows a comparison that we have made between EEP predictions of the housing energy use for the Neath Port Talbot region, and actual gas fuel use in the area (as determined from BERR data[1]). Each point on the graph is for a "middle layer super-output area"; each contains roughly 3500 dwellings of mixed type and age. This agreement was reached using the fully realised EEP model resulting from the manual survey.

The manual survey method used for this region provided a level of information that was otherwise unobtainable. However an investment in manpower and time on this scale has proved to be a barrier to the wider uptake of the EEP model, or indeed others like it. In order to allow greater access to such modelling methods, there is a need to explore and develop more efficient methods for acquiring building stock information. It is expected that a more efficient survey method would provide similar quality results for a lower investment. Should such improvements become available, feedback from EEP users indicates that detailed regional energy models would find wide-scale application in local and regional government.

BUILT FORM INDICATORS

Experience from Neath Port Talbot and similar surveys led us to believe that (in a UK context at any rate) building era could be determined by eye; that is a Victorian building is visually distinguishable from interwar housing and from 1960's and 70's housing. As illustrated in figure 2, architectural and planning "styles" provide convenient visual clues for identification. Development in the UK has followed certain broad patterns that allow for the identification of buildings by their plan form. For example, housing built prior to WWI is distinguishable by the back-wing form – a main building with rear extension, usually containing the kitchen and scullery. In the inter-war period, a defining feature was the rise in popularity of the semi-detached home. Not only did this form provide rear access without the need for back lanes (Edwards 1981), but they also provided a feeling of space and openness in reaction against bye-law housing. Housing from this later period also often shows a degree of uniformity, largely due to legislative directions aimed at improving substandard housing (Barrett and Phillips 1987; Colquhoun 1999).

An initial subjective comparison of map data and survey data acquired for the Neath Port Talbot region lead to the hypotheses a) that housing of a common era exhibit common plan forms and b) that these common forms alter with the era of construction. Geometrical analysis of building plan forms should therefore be able to distinguish approximate building age. For example, figure 3 shows typical plan forms of two ages of housing, extracted from the region studied. Although the two forms are similar, the pre-1919 houses exhibit a consistently higher aspect ratio and so should be robustly identifiable. Other similar distinctions were observed between other housing eras.

At the time of this initial investigation, digitally encoded maps were becoming readily available. These contained, in addition to road layout, representations of plot and building footprints in polygon form. Within a GIS system, these map polygons could be identified and geometrically analysed, enabling a potentially automated analysis method. The questions were raised: could this digital information be used to aid in the automatic identification of building age? ; could the qualitative differences in housing eras be quantified?

METHODOLOGY

The concept was tested by devising methods of analysis of digital map data and comparing their ability to identify built age in the region which had previously been manually surveyed in detail. The power of the analysis methods could be tested against samples taken from this database to establish applicability and potential accuracy. The region contains a diverse mix of stock, as shown in table 1, and so should provide a reasonable indication of overall identification power in a UK context.

Two different approaches to map analysis were investigated; 1) Shape: an analysis based on the footprint of the buildings, as described in the preceding section, and 2) Context: an analysis based on information about each building's nearby surroundings, such as distance to road.

Shape based methods

Shape methods quantify the outline of the building footprint as extracted from the map. Apart from rectangles, buildings can rarely be expected to take on classical polygon forms, so methods to encode, compare, and group general polygons are required. Two methods for this have been explored; Raster coding and Vertex coding.

Raster coding

Raster coding involves overlaying an analysis grid to encode the vertices of the polygon. This creates a code "number" for the shape. Similar building shapes should define the same code, and so the code

[1] http://www.berr.gov.uk/whatwedo/energy/statistics/regional/index.html

provides the shape clustering automatically. A relatively coarse grid provides some tolerance to map "errors" e.g. a slightly misplaced vertex.

In the GIS system a fixed grid is overlaid on a normalized (i.e. making the longest length unitary) building plan outline, with the building centered in the grid. Internal GIS functions are used to locate nodes (corners) of the building shape; where a node lies in a grid cell this is marked as occupied otherwise the cell is marked empty. As illustrated in figure 4, the occupied/empty cells are then encoded to form a number; in this case 010001110000100001. The resulting codes can be checked for mirroring and rotations.

Vertex coding

Vertex coding utilises a chain coding approach (as described in Fu 1982) to encode the shape of the building plan outline as a sequence of seqment lengths, where each segment, or wall, denotes a significant change of direction. The GIS system can provide the location of vertices and segment lengths for any selected polygon. The information derived from the outline is the number of segments, the total perimeter length and the chain code of the segments, as illustrated in figure 5. Each code is potentially unique, and further processing is required to group similar chains to form shape clusters. As with the Raster method, the codes are first normalised for longest length before further comparison.

Small differences in form, due perhaps to map drawing errors or insignificant differences in shape, can produce unique chain-codes. A method that produces as many clusters as buildings will be unuseful; the method must group together *similar* shapes. In order to group similar building plan forms, a tolerance to differences must be imposed. Any two normalised outlines (of polygons with the same number of vertices) can be compared vertex to vertex and a difference, or variance, parameter ∇ between any two plan forms can be defined as; $\nabla = \sum_{i=1}^{n}\left(V_i a - V_i b\right)^2$, where V is a vertex position for plans a and b with n vertices. In order to allow for rotation and mirroring, ∇ can be minimised for any pair. The variance parameter ∇ is used to aid in the formation of useful (i.e. populated) clusters, by defining a difference tolerance $\varepsilon\nabla$?such that a cluster contains shapes with a low difference; $\nabla \pm \varepsilon\nabla$. This is illustrated in figure 6, where the tolerance can be set to distinguish between similar and different shapes; with a tolerance of 4 only the first two candidates cluster with the prototype. In our method, each polygon representing a building in a sample can be tested and either form a new group if it is "close" to no others, or be placed in an existing, closest, group. In a two pass exercise, the members of each group are then used to form an average or "prototype" for the group, and the whole population once again tested against these prototypes to form the final clusters of plan shapes.

In these methods, the quality of the grouping will obviously vary considerably according to the parameters used; for instance in the raster method, with the grid size, or in the vertex method, the difference tolerance. In the extremes, two obvious degenerate results exist; each building is in a unique group (tolerance too low) or all buildings are in one group (tolerance too high). A useful result will be a moderate number of groups, with each group containing a large number, and only a few unique (or solo member) groups.

The testing of these shape methods will focus on two questions; can useful (e.g. highly populated) clusters be formed, and once formed, can those clusters inform of built age?

Context based method

The built form reflects the complex political, economic, social, and cultural processes in place at the time of construction, and these processes, of course, change with time. Another approach to establishing built age from map data is therefore to move the focus from the footprint to the building's surroundings; its context.

As with footprint shape, the history of housing development leads to identifiable patterns in the characteristics of the building's relationship to its surroundings. Early 20th century housing favoured the terrace, while later construction is almost exclusively detached or semi-detached homes. These are reflected in the geometrical nature of the property on which the building is placed (the plot) and in the distances between neighbouring buildings. Barr and Barnsley (2004), for instance, used maps to infer urban land use and successfully identified areas with similar built ages by considering street layout patterns. Our intent is to extend this type of analysis to identify the age of individual properties, which may or may not be of similar age to the surrounding development.

For our context analysis, factors were derived based on:- from past experience; their subjective correlation with the built age; the results of a built morphology review; and, mainly for pre-war and inter-war housing, an existing literature. Only indicators that could be readily obtained through geometrical analysis of digital maps were considered. We have devised five such building form indicators:

1. *Plot size*; the total area of the land associated with the building;
2. *Distance to road*; the length from the building main entrance to the nearest roadway;
3. *Road angle*; the angle made from the building frontage to the nearest roadway;
4. *Number linked to*; the number of other buildings directly touching the target

building, for instance a detached house has a link of 0, while a semi-detached has a value of 1 and a mid-terrace a value of 2;

5. *Terrace length*; the total number of buildings that are directly or indirectly linked to the target.

These five parameters were derived for all buildings in the sample, via the GIS system. That data was then used to form a multinomial logistic regression against built age, using the software SPSS. If statistically significant relationships are observed, then age can be inferred from the context form.

RESULTS

Our development and testing of the methods was based on the Ordnance Survey's MasterMap™ product, which has a layer of polygon data that is grouped as regions or objects. These objects have many items of data associated with them; two such items are the "Descriptive Group" and the "Descriptive Term". These two describe layers of 1) buildings, 2) roads and paths and 3) plots within which the buildings sit. For analysis, these maps are embedded within a GIS system; Mapinfo™ in our case. Each of the three methods described above were implemented in Mapbasic™, and tested against purposefully selected samples of the Neath Port Talbot survey area, with each sample containing several thousands of dwellings. The samples could not be completely randomised as: a) we wanted a wide range of ages to be represented in each, and b) the context method requires contiguous areas rather than individual buildings.

Raster coding

As expected, grid density had a significant impact on this method's ability to form useful (e.g. highly populated) clusters. Grids of 6x3, 8x4, and 8x8 were examined, and of these the 8x4 grid produced the best results. Success rates dropped considerably for the others; the grid was either too coarse to delineate difference in shape, or conversely too detailed so that unimportant or erroneous details were included in the codes being generated.

The 8x4 grid produced 491 clusters from a sample of 2000 dwellings. However many of these were "solo" forms; that is the cluster contained only one building. Excluding those, 132 populated clusters were discovered by the method; of the 132, 83 clusters contained a unique building era, while the others contained a mix of eras. Sample plan form "clusters" resulting from this analysis are shown in figure 7. The best agreement (table 2) was found for pre-1919 housing; in this era 69% of the stock in the sample was attributed to unique plan forms (that is, forms not associated with other eras). There were similar clusters found for other ages but success rates were lower; for instance 36% of post-1980 housing and 26% of 1945-1965 housing were correctly identified, however only 12% of 1965-1980 housing were attributable to unique plan forms by this method.

Vertex coding

A more detailed inspection of the map polygons revealed a number of challenges to implementation of this method. Not all buildings were drawn consistently or accurately; for instance extra, redundant, nodes were often placed within a line segment. While this would not affect the original intended use of these maps, they could affect our analyses. Many of these issues were resolved by introducing a tolerance factor; nodes that were within a smallest significance distance (e.g. 0.5m) from another, or represented an angular change of less than 15 degree, were ignored. This would lead to the loss of small details in built form, but this was not considered a significant loss for our purposes.

In a sample of 7144 buildings, 2471 clusters of 2 or more buildings were created, of these 55 were "large" clusters (of 10 or more members). 1390 "solo" clusters were found.

Although the number of clusters produced were high, as shown in table 2 there is an improvement in the identification rates of crucial building eras; pre-1919, and 1919-1945, over the raster coding approach.

Context coding

Implementation of this method showed that the derivation of some of our indicators from the existing map data were problematic. For example, as building frontages or entrances were not explicitly identified in the map data or objects, the frontage of a building could only be inferred to be that closest to the nearest road.

Determining, from the map data, plot area for a property also proved less than obvious. For instance, within the map, a front and back garden of a terrace property may be stored as two polygons objects. In addition, a plot of land may touch more that one property, so that it may appear to "belong" to more than one building. In our analysis, property plots were defined as any land plot attached to the building in question. This simplification results in a proxy for urban density, but may lead to double counting in certain circumstances. In these cases, errors in analysis may be introduced by our assumptions or definitions.

The statistical analysis of the five parameters previously described, using a sample of 7000 dwellings, showed that the best-fit model could correctly classify over 76% of built form over all eras. The pseudo-R statistics, used to determine the proportion of variation explained by the regression model, are 0.636 for the Cox and Snell statistic and 0.696 for the Nagelkerke statistic; these indicate that the regression model was fairly good in accounting for the differences in classification.

As with the other methods, the pre-1919 era buildings showed a high success rate for identification, however the Context method was also very successful in identifying the 1945-1965 era, with a 95% success rate for those properties. However, the classification accuracy of other age periods varied significantly (see table 2). The identification of modern era housing was notable; only 1% was considered by this method to have unique forms. In fact, most modern dwellings were placed erroneously into the 1945-1964 era.

Depending on the age of the building, certain of our indicators were found to be more important than others. For example, for homes built before 1919, the distance to road (sig. 0.001) was one of the most significant indicators, presumably because pre-war terraces were almost uniformly sited very close to or on the front property line. In other age groups, most notably homes built after 1965, there is too much variation in this descriptor for it to be statistically significant.

DISCUSSION OF RESULTS

Table 2 compares the ability of the various methods to resolve building age within the samples tested.

We feel that the results indicate that the analysis methods are capable of analysing digital map data and producing indicators of building age. Building plan forms do cluster and there are correlations with age. However we do not consider the methods sufficiently accurate for general use; further testing and development is required. Of those methods tested, the Vertex code method and the Context method appear to be the most promising for further development.

Identification of pre-1919 housing appears to be relatively straightforward, as both the shape and the context methods were capable of producing identification rates over 90% for this era. This is considered highly encouraging as this era marks a crucial housing type often characterised by solid wall constructions, and a correspondingly poor energy performance; the “hard to heat” home.

On the other hand, the following era; 1919-1945, also typically shows relatively poor energy efficiency, yet the methods have been less successful in identifying buildings in that era. The Context method appears to be more powerful in identifying that era, yet is less powerful that the Vertex method in others. There would therefore appear to be benefit in combining the two methods. At the time of writing, the two methods were implemented and tested independently of each other, and we are now exploring methods to couple the context and vertex approaches.

Mapping and inspection of properties that were mis-identified by the methods has provided several interesting results. Firstly, it appears that in a number of cases the models were actually correctly identifying the built age; it was the initial manual survey that was erroneous and a second visit prompted by this finding showed many anomalies. Secondly, mis-identified properties were often placed only one era off. For example, if the most current eras are combined into one category spanning 1965 to the present, the identification success rate is over 90%. However, such a wide age band would not be useful in determining thermal properties of the buildings. Given the often gradual transition in building construction styles, this uncertainty in identification is understandable, but it implies that there will be a maximum accuracy to which built age can be expected to be identified by an automatic analysis.

Calculations made using a simple error propagation model suggests that, in a population of buildings with a spread of age characteristics representative of the UK, the accuracy of the estimate of the total energy use of the population can withstand even significant errors in age identification. Since the average energy requirements of different building ages are relatively similar (e.g., from table 1, pre-1919 = 39.5 MWh/year, compared to post 1980 = 25.8 MWh/year), and since the total number of properties is accurately known (so that an error in identification places a property in another era), even 30% uncertainty in identification leads to only 3% uncertainty in the total regional energy use.

A more demanding, and potentially more interesting, use of this data/model is to estimate the energy savings accruing from an initiative or investment; for instance, grants to improve the uptake of cavity wall insulation. An acceptable prediction of energy savings over the region (±10% for instance[2]), would, for our test region, require an accurate (to 90%) estimate of pre-1919 stock (in order to remove them from the calculation) and of 1945-1964 stock (as the most prevalent poorly performing era). The other, more efficient stock has less impact on the results, so that even identification accuracies as low as 30% for those eras has little effect. This leads us to believe that we are approaching a useful method, in particular if we can combine the two approaches and successfully identify the eras from pe-1919 through to 1945-1964.

We have yet to test these methods on the full data set of 55,000 dwellings to establish ultimate accuracy for the region. In general, the initial implementations of the methods in MapBasic were notably inefficient and could require days for analysis of the samples used here. Recent improvements of the algorithms and coding used have considerably improved performance; calculation times are now in the order of minutes, and so full regional analyses are now being undertaken.

We now consider it unlikely that any method will be able to fully identify the age of buildings in a region; there will be an upper limit of the accuracy of

[2] Note that this accuracy refers to that resulting from the identification of stock only, not to the accuracy of the underlying energy model.

identification. Changes in built "style" are notable but gradual, and so there will naturally be many properties that appear "out of time" as they straddle the eras. The digital maps that form the basis for the methods are not error free, and even the labour intensive "walk-by" method introduced notable errors in identification. It is likely therefore that each property can only have an age probability determined by any such analysis, providing an age probability distribution (for illustration, for a particular property, a 2% chance it is pre-1919; 4% 1919-1945; 75% 1945-1964; 19% 1965-1980; and 1% post 1980). Calculation methods for regional building energy use should be adapted to recognise such uncertainty in classification, and to propagate these uncertainties to the final energy result. Once implemented however, such a statistical approach would also allow uncertainty to be placed on other parameters, for instance probabilities for retrofit double-glazing, conservatories, or loft rooms could be attached to each era. In order to use and present this information, we consider that the EEP model, for instance, must become more stochastically based, as opposed to the deterministic approach used at present. This avenue is currently being explored.

Finally, when the analyses were undertaken, many interesting anomalies occurred in the results, particularly in the context method. On inspection, these anomalies in context were related to mapping errors (for instance unclosed polygons). This leads us to believe that the techniques developed here may also have a more general utility in discovering mapping errors during the construction or revision of urban maps.

CONCLUSIONS

A number of potential methods for the automated analysis of digital maps have been tested. It has been shown that building age, at least in terms of era, can be inferred from a geometrical analysis of footprint shape or building context. However, while pre-1919 housing can be readily identified, with accuracies greater than 90%, other equally important eras are less well identified. The results are considered encouraging, but further development and combination of the recognition methods is required to bring the overall accuracy up to a generally useful level. In particular, further development is required to combine the two most promising approaches.

Results so far are specific to the region studied; while the authors believe the region studied is not atypical of British suburban housing, we make no suggestion that the predictive models so far generated will work with other countries, or indeed other towns. In application to a "new" region, the models will need to be constructed. It is expected that reasonable accuracy will be attained by "training" the system on a sample of the population to be modelled; while this sample must be manually surveyed to determine the building stock characteristics (age, construction type etc.), the size of the sample should be such that the effort required is significantly reduced, so that the modelling effort becomes more attractive and economic. If a region can be satisfactorily described by a sample of, for illustration 20%, this implies a significant reduction in time and effort required for the survey; 3-4 man-months investment rather than 18.

An automatic age identification method, even if not perfect, will allow a much more efficient survey method, where manual inspection can be targeted to areas or specific buildings where uncertainty exists.

It is likely that there will be an upper limit of the accuracy of identification of built age; even the database produced by a manual "walk-by" method has been found to contain significant errors in identification. This lead us to believe that the next generation of regional models such as EEP must embrace this uncertainty and include it in its calculations, leading to a more stochastic, rather than deterministic, energy calculation methodology.

ACKNOWLEDGEMENT

This work, and the initial development of the original Energy and Environment Prediction (EEP) model, has been funded by the Engineering and Physical Sciences Research Council. The authors thank Neath Port Talbot Borough Council for their support.

REFERENCES

Barr, L.S., Barnsley, M.J., and Steel, A., 2004, On the separability of urban land-use categories in fine spatial scale land-cover data using structural pattern recognition, Environment and Planning B: Planning and Design, volume 31(3) May, pages 397 – 418

Barrett, H. and Phillips, J. 1987. Suburban Style: The British Home, 1840-1960. London: McDonald & Co.

Boardman, B., Darby, S., Killip, G., et al., 2005, 40% House Final Report, Environmental Change Institute, ISBN 1 874370 39 7

Colquhoun, I. 1999. RIBA Book of 20th Century British Housing. Oxford: Butterworth-Heinemann

Edwards, A. M. 1981. The Design of Suburbia: A Critical Study in Environmental History. London Pembridge Press.

Fu, K.S. 1982. Syntactic pattern recognition and applications, Prentice-Hall

Jones, P., Williams, J.L., and Lannon, S., 2000, Planning for a Sustainable City: an Energy and Environmental Prediction Model, Journal of Environmental Planning and Management, vol. 43, no. 6, pp. 855-872(18)

Building era	*Region component*	*Average type fuel use MWh/year*
Pre 1919	35%	39.5
1919 – 1944	13%	38.1
1945 – 1964	27%	29.9
1965 – 1980	18%	32.1
Post 1980	7%	25.8

Table 1 : Stock profile of the Neath Port Talbot survey region; 55,000 dwellings in total.

	Method:		
Building era:	Raster	Vertex	Context
Pre 1919	69%	94%	93%
1919 – 1944	28%	70%	32%
1945 – 1964	26%	20%	95%
1965 – 1980	12%	66%	19%
Post 1980	36%	55%	1%

Table 2 : Proportion of successful age identification for each method.

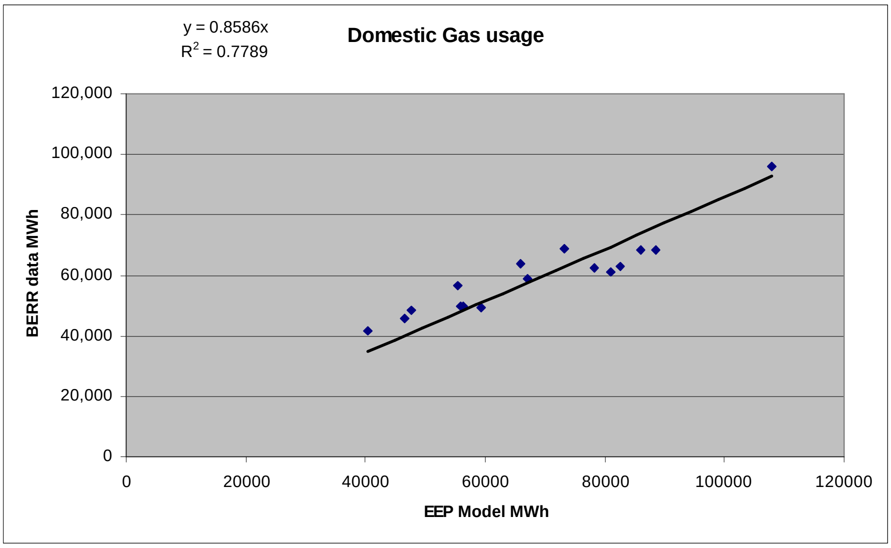
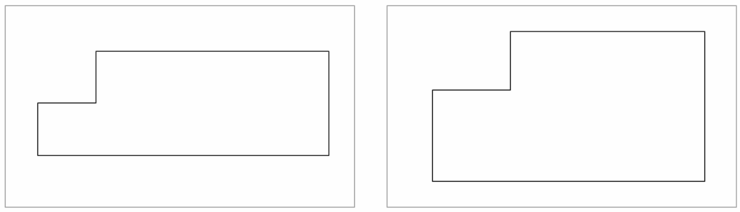


Figure 1: Comparison of regional energy model and actual fuel use.

Figure 2 Two eras in terrace housing.



a) Pre 1919 housing *b) 1919 to 1945 housing*

Figure 3: Typical dwelling outline patterns.

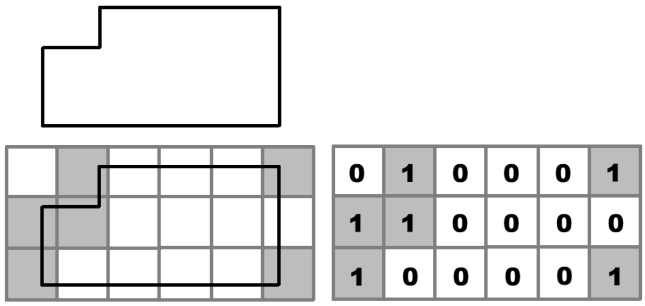


Figure 4: 6x3 raster coding for a building outline.

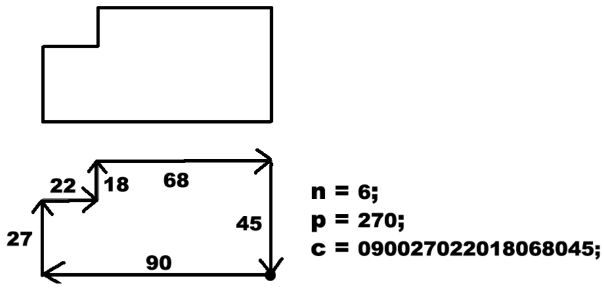


Figure 5. Vertex coding, in dm, for a building outline.

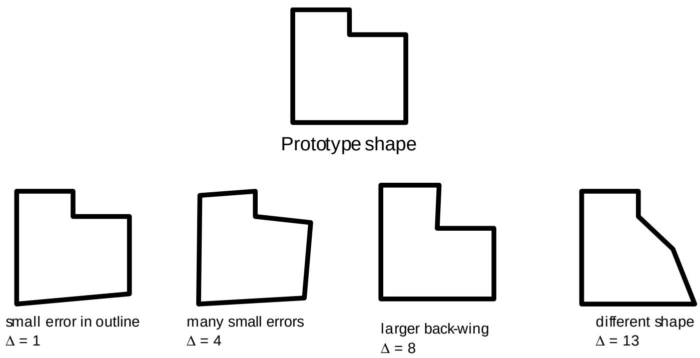
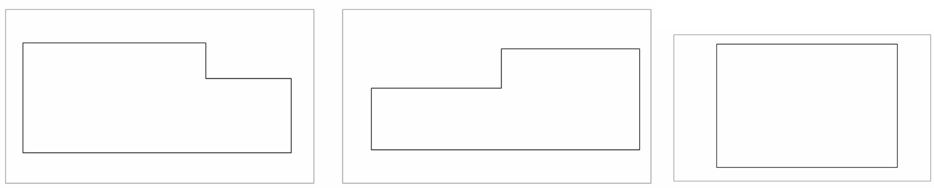


Figure 6. Differences in Difference factor D between similar plan forms.



a) *b)* *c)*

Figure 7. Example plan forms generated by the raster method; a and b uniquely define an era, while c) covers a number of eras.