Application of clustering technique for the segmentation of building stocks to renovate

Yunseok Lee, Pierre Boisson, Mathieu Rivallain, Olivier Baverel

Application de techniques de clustering pour la segmentation de parcs de bâtiments à rénover

Yunseok LEE *[1], Pierre BOISSON[1], Mathieu RIVALLAIN[1], Olivier BAVEREL[2]

[1] Université Paris-Est, CSTB
84 Avenue Jean Jaurès, 77420 Champs-sur-Marne,
[2] Université Paris-Est, Laboratoire Navier (UMR 8205), ENPC, IFSTTAR, CNRS
Cité Descartes, 6-8 Avenue Blaise Pascal, 77455 Champs-sur-Marne,
***yunseok.lee@cstb.fr**

RESUME. L'élaboration de stratégies de rénovation énergétique à l'échelle d'un parc existant nécessite de modéliser l'ensemble des bâtiments, en tenant compte de l'information disponible. Les approches de modélisation recourent traditionnellement à des typologies, régulièrement fondées sur les caractéristiques constructives connues ou observables des bâtiments. Or il s'avère difficile d'établir des corrélations entre les performances énergétiques et ces typologies descriptives. L'objectif de cette contribution vise à développer une approche de segmentation du parc fondée sur des techniques d'apprentissage automatique. Différents algorithmes de clustering ont été appliqués sur une base de données réelle de logements français. Leur paramétrisation et le choix du nombre de clusters, ont été étudiés en vue d'apprécier la pertinence de cette approche. Cette première application donne des résultats encourageants pour segmenter un parc de bâtiments selon des critères descriptifs mais aussi performanciels (consommations ou économies d'énergie, couts d'investissement liés à des scénarios de réhabilitation).

MOTS-CLÉS : parcs de bâtiments, clustering, performance énergétique

ABSTRACT. The development of energy renovation strategies at the scale of an existing stock requires to model all buildings, taking into account the available information. Modeling approaches traditionally employ typologies regularly based on observable or known construction characteristics of buildings. But it is difficult to establish correlations between energy performance and these descriptive typologies. The aim of this contribution is to develop a building stock segmentation approach based on machine learning techniques. Different clustering algorithms have been applied to a real database of French homes. Their parameterization and the selection of the number of clusters were studied in order to assess the relevance of this approach. This first application gives encouraging results to segment a building stock according to descriptive but also performance criteria (such as energy consumption, energy savings and investment costs related to refurbishment scenarios).

KEYWORDS : building stock, clustering, energy performance

1. INTRODUCTION

1.1. TYPOLOGY OF EXISTING BUILDING STOCKS

The renovation of the existing stock is the main source of energy savings in the building sector in France. At the scale of a stock, the development of retrofit strategies need to model all of the buildings, considering the available information. Modeling approaches traditionally employ descriptive typologies to define the criteria of typical buildings and describe exemplary buildings of representing building types (TABULA Project Team 2012). The exemplary building, or the reference building might be "a hypothetical or real reference building that represents the typical building geometry, components and

systems, typical energy performance for both building envelope and systems, typical functionality and typical cost structure in the Member State and is representative of climatic conditions and geographic location" (European Commission 2012).

From the state of art, many different projects exploiting typologies of building stocks are found. While the typologies are defined by different sets of specific attributes depending on the issue considered, most of the attributes are observable features used for conventional urban or architectural issues. However, these attributes cannot guarantee appropriate building typologies on expected energy performance issues. For example, the period of construction and the living surface, quite common attributes in the state of art, cannot show sufficient discrimination in between similar buildings after different process of renovation over time.

Ultimately, from the perspective of energy performance, successful building stock modeling requires more than the traditional building typologies. To the expectation for robust building stock modeling, machine learning techniques might offer powerful tools to analyze wide sets of buildings, and identify potential hidden structures of building stocks.

1.2. MACHINE LEARNING

Machine Learning is a broad field of study in computer science mainly concerned with the discovery of regularities, such as models and patterns, in data (Fürnkranz 2012). Machine learning is roughly classified into two main approaches depending on the nature of the learning signal. **Supervised learning** exploits a general rule that maps inputs to outputs, from labeled training data, given sets of example inputs and their desired outputs, e.g. automatic spam filters of e-mail providers. **Unsupervised learning** searches for some intrinsic structure in the example inputs. Unlike the supervised learning, the data is simply an input without an associated output, e.g. grouping of similar news articles. Among unsupervised learning techniques, data clustering is a method of grouping a given input sets based on their similarities. Consequently, data clustering is to reduce a number of examples into smaller number of clusters, giving a clue for building stock modeling.

2. METHODOLOGY

2.1. DATA OF EXISTING BUILDING STOCK

In this study, we used the PHEBUS survey data, which was executed by INSEE, the French national statistics bureau, in 2013 for the purpose of providing energy performance data of French residential building stock. The PHEBUS survey consists of two frames; the first frame CLODE dealt with general properties, socio-demographic characteristics and energy behavior through face-to-face interview with 8000 representative houses. In the second frame DPE, which is more relevant to this study, the energy performances of housing were measured by qualified diagnosticians.

The PHEBUS DPE data presented detailed information about housing descriptions, energy consumptions, and building information, which includes opaque walls, windows, thermal bridges, heating systems and domestic hot water systems. Depending on the diagnosis method and the type of housing, some information was not recorded in the database. For example, the houses examined with energy bills have only energy meter reading with basic building information. On the other hand, the housings estimated with 3CL-DPE, a conventional method to estimate energy consumption, have more

detailed information of the building, lacking the actual metered data. Some houses have both data, giving different actual and calculated energy consumption.

The PHEBUS DPE data provides interesting comparison among segments of conventional typologies. When the final energy consumption for heating per unit surface of houses is focused on, collective houses consume less energy than individual houses. Though recently built houses tend to consume less energy as well, the houses built before 1980 have no apparent tendency depending on the period of construction. The houses using electricity as the only energy source show distinct low heating energy consumption than those using combustible sources (Figure 1). Some features can be calculated or obtained through combinations with other data, such as heating degree day depending on department and elevation.

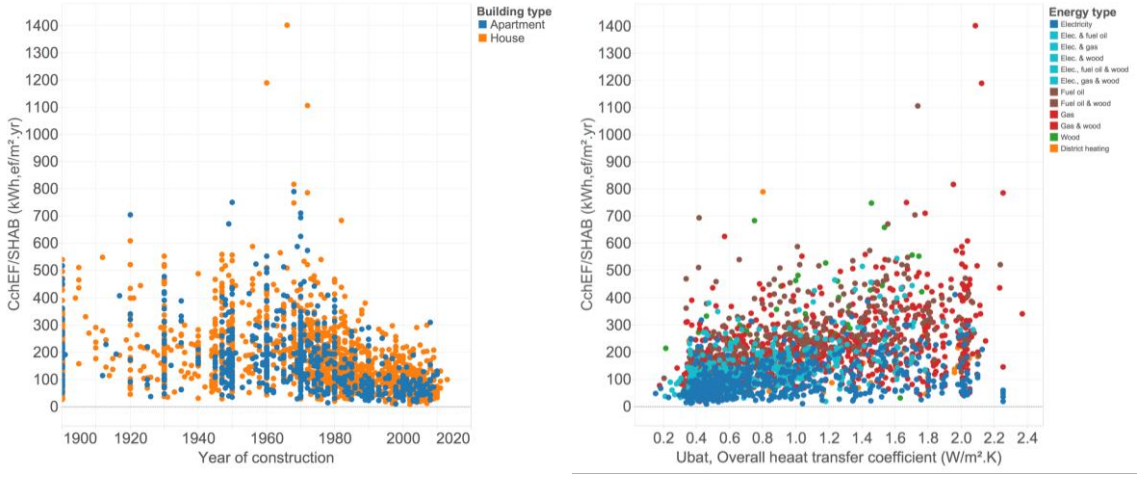


Figure 1 : Final energy consumption for heating according to some features.

2.2. ATTRIBUTES SELECTION

More than 240 features are included in PHEBUS DPE data. In addition to these, many other interesting features can be newly added by recombination of the existing data or combination with external data. Among the features, however, not every feature is meaningful for a specific purpose and some features are often redundant. Before a study, therefore, appropriate features should be selected among them, and the features to be added should be decided. In this study, for this purpose, a dozen features in building information and heating system, which seem more relevant to the energy performance of housing, were chosen. Though some of them were directly available from the PHEBUS DPE data file, others should be calculated combining with other internal (e.g. window area) or external data of the file (e.g. heating degree-hour according to the department).

Features	Minimum	Maximum	Average
Living space (SHAB, m^2)	15	2005	109.20
Heating degree-hour (Dh) corrected for elevation (°C.h)	31333.8	106568.2	58757.2
Heat loss due to air changes (DR) per SHAB ($W/m^2.K$)	0.25	0.93	0.63
Overall heat transfer coefficient (U_{bat}, $W/m^2.K$)	0.15	2.37	0.99
Window-floor surface ratio	0.00	0.52	0.15
Compactness	0.08	12.00	2.48
Final energy consumption for heating ($kWh_{FE}/m^2.yr$)	8.92	1400.62	176.61
Thermal efficiency of heating system	0.50	3.61	0.86

(a) Numeric features

Features	Number of categories	Remark
Type of building	2	Apartment & individual house
Period of construction	11	Ordered (‘until 1850’ to ‘after 2012’)
Thermal mass class	4	Ordered (‘light’ to ‘very heavy’)
Type of energy	13	Combination of 6 energies

(b) Categorical features

Table 1 : Selected features in PHEBUS DPE data.

Redundant features can degrade the performance of analysis. To eliminate the redundancy, the linear association between features can be considered before the analysis. The Pearson’s product-moment coefficient ρ can be calculated between two numeric features X, Y as following equation,

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad \textbf{(1)}$$

where, cov , E, σ and μ stand for covariance, expectation, standard deviation and mean value respectively. Pearson suggested an interpretation of the coefficient size (Table 2) (Pearson 1904).

As the Pearson’s correlation coefficient can be calculated between numeric features, in this study, the 8 numeric features and a categorical feature expressed in numeric form are applicable. A matrix plot of correlation coefficients shows that final energy consumption for heating has moderate correlations with overall heat transfer coefficient, heating-degree-hour corrected for elevation, thermal efficiency of heating system, and compactness (Figure 2).

Size of correlation	Proposed interpretation
0.75-1.00	High correlation
0.50-0.75	Considerable correlation
0.25-0.50	Moderate correlation
0.00-0.25	Low correlation

Table 2 : Interpretation of a correlation coefficient

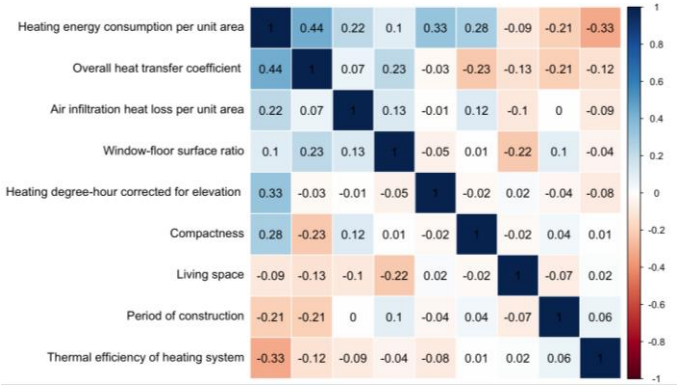


Figure 2 : Correlation matrix plot of numeric features.

The multiple criteria of decision making problems can be represented with multiple objective functions. The vector space defined with objective variables of the multiple objective functions is called **objective space**. On the other hand, the vector space with decision variables is called **decision space**. The vector space consisted of both variables can be considered as a **hybrid space**. In this study, considering the energy performance, the final energy consumption for heating forms an objective vector, thereby objective space. The others compose decision variables and decision space (Table 3).

Objective space	Decision space
Final energy consumption for heating	Living space (SHAB), heating degree-hour (Dh) corrected for elevation, heat loss due to air changes (DR) per SHAB, overall heat transfer coefficient (U_{bat}), window-floor surface ratio, compactness, thermal efficiency of heating system, type of building, period of construction, thermal mass class, and type of energy

Table 3 : Classification of features by space.

2.3. PREPROCESS OF DATA

The PHEBUS DPE data contains 3452 items, which include entire building data, incompletions and errors. Duplicated items, of which energy consumptions were calculated in 3CL-DPE method and collected from energy bills in parallel, exist as well. These apparently unnecessary data could degrade the performance of clustering analysis by distorting results. Therefore, eliminating building data, overlapped data, and incomplete data and unreasonable data, 2339 houses were finally left to be used in the analysis.

Based on the distance conception, most of machine learning algorithms can handle only numeric variables. Categorical variables should be converted into numeric variables by introducing dummy variables for each category. If a categorical feature has an order, the values can rather be expressed with ordinal numbers. While the features expressed in the form of numbers are available for machine learning, the results ca be influenced by the scales of features. This scale dependency can be avoided by standardization or normalization of the numeric values.

In this study, two categorical features (type of building and type of energy) were converted into binary variables, two others (period of construction and thermal mass class) into normalized ordinal numbers. The eight numeric features were standardized to have zero mean and unit variance.

2.4. APPLICATION OF CLUSTERING ALGORITHMS

Among a variety of clustering algorithms, the following five algorithms were selected in this study. **K-means** is a flat clustering algorithm. The objective of K-means algorithm is to find a set of clusters which minimize the residual sum of squares (RSS) (Hartigan 1979). **Hierarchical Agglomerate clustering (HAC)** is a connectivity-based cluster analysis algorithm, which builds a hierarchy of clusters by a bottom-up approach (Murtagh 1983). In order to increase the quality of clusters, some algorithms integrated hierarchical clustering and distance-based clustering. Among the algorithms, **BIRCH** algorithm introduces a CF (Clustering Features) concept for effective clustering particularly over large data-sets (Zhang et al. 1997). **Affinity propagation (AP)** is an exemplar-based clustering algorithm. In this algorithm, data points exchange *messages* to find their *exemplars*, and the group of data points which shares the same exemplar becomes a cluster (Dueck 2009). **DBSCAN** is a density-based clustering algorithm which can handle arbitrary shape clusters. This algorithm requires two parameters defining the least dense cluster (Ester et al. 1996).

2.5. EVALUATION OF CLUSTERING PERFORMANCE

Among various evaluating indices of the clustering performance, the silhouette coefficient is widely used when the ground truth classes are unknown. The silhouette coefficient was suggested to be used to select the number of clusters for partitioning techniques (Rousseeuw 1987). For each object i, when $a(i)$ is the average dissimilarity of i to all other objects of the belonging cluster A, and $b(i)$ is the average dissimilarity of i to all other objects of the second-best cluster B, the silhouette value $s(i)$ is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \textbf{(2)}$$

The average of the $s(i)$ over all object i is called the overall average silhouette width, and the highest average silhouette width over all the number of clusters is called the silhouette coefficient. A subjective interpretation was proposed from the experiences (Table 4).

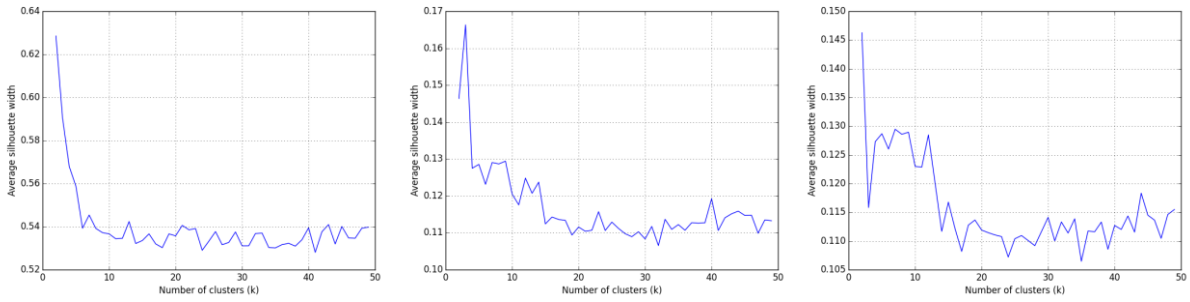
Silhouette coefficient	Proposed interpretation
0.71-1.00	A strong structure has been found.
0.51-0.70	A reasonable structure has been found.
0.26-0.50	The structure is weak and could be artificial, try additional methods.
≤0.25	No substantial structure has been found.

Table 4 : Interpretation of the silhouette coefficient (Struyf et al. 1996).

3. RESULTS AND DISCUSSION

3.1. DECISION OF NUMBER OF CLUSTERS

Some algorithms introduced in 2.4, typically K-means, require the number of clusters as an input. As many real-world problems, however, it is unknown in this study. Thus, the silhouette values were calculated for a range of numbers of clusters in each space (Figure 3). The objective space, which has the smallest one dimension, showed reasonable structures with the average silhouette width around 0.54. No substantial structures could be found in the decision space (15-dimensional) and the hybrid space (16-dimensional). Instead of two to five clusters which seemed not to be sufficient for the partitioning of building stocks, seven were selected as the number of clusters. This number of clusters would be applied in the decision and the hybrid space as well.



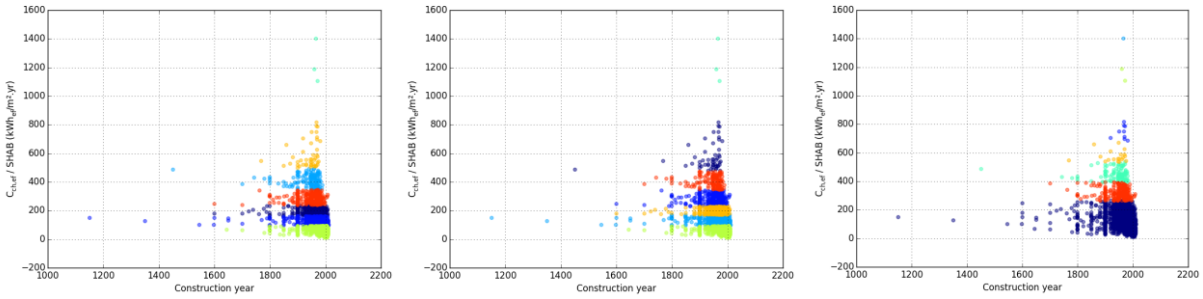
(a) Objective space *(b) Decision space* *(c) Hybrid space*

Figure 3 : Average silhouette width with different number of clusters

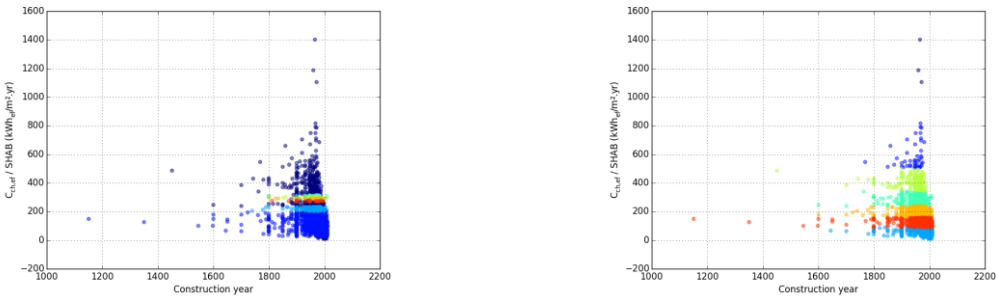
3.2. COMPARISON OF ALGORITHMS

Five algorithms mentioned in 2.4 were applied to the prepared data. The number of clusters was set to 7 for three algorithms, K-means, HAC and BIRCH, where the number of clusters could be selected. For the others, varying available parameters, as close to 7 clusters as possible were obtained.

The characteristics of the algorithms could be well observed in the objective space, the simplest space. While three algorithms, i.e. K-means, HAC and AP formed similar clusters, BIRCH and DBSCAN showed quite different distributions of clusters. Particularly, DBSCAN, the only algorithm allowing noise points, which do not belong to any cluster, categorized from 13.9% (in the objective space) up to 31.8% (in the hybrid space) of houses as noise. As DBSCAN tends to equal density clusters, in the data without apparent density distinction, such as in this study, the clustering performance seems to decline. On the other hand, though HAC and BIRCH share some theoretical background, the results were quite different. In the case of AP algorithm, where the number of clusters could be achieved by trial and error, the result was similar to those of K-means and HAC, the two most conventional algorithms.



(a) K-means *(b) HAC* *(c) BIRCH*

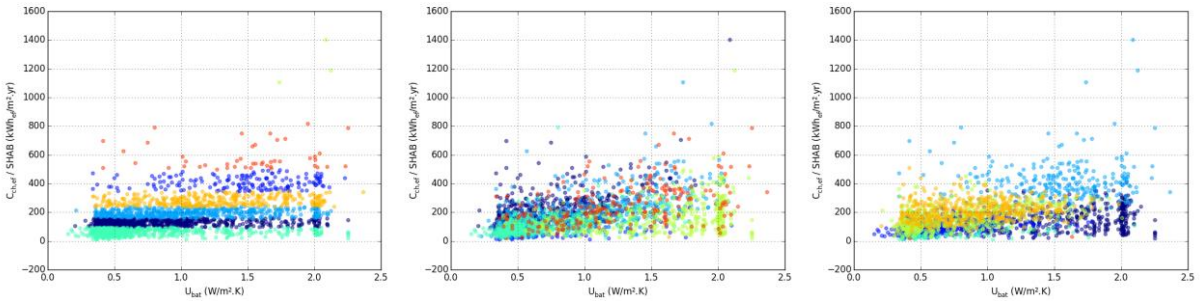


(d) DBSCAN *(e) AP*

Figure 4 : Algorithm comparisons in the objective space (k=7).

3.3. COMPARISON OF SPACES

Clustering analysis were performed in three different criteria spaces, respectively. Though the decision space and the hybrid space can be subdivided by the including features, to simplify the problem, the spaces were presumed to include all 11 and 12 features, respectively. In this section, as a representative case, the results of K-means algorithms were presented.



(a) Objective space *(b) Decision space* *(c) Hybrid space*

Figure 5 : Space comparison in objective space (K-means, k=7).

Clustering in objective space did not show distinguishable clusters in the decision space (x-axis in *Figure 5 (a)*) and vice versa (y-axis in *Figure 5 (b)*). As a feature occupied just one dimension in the decision space, the clusters were not evident even in an aspect of the decision space. It is worthy of notice that, in the hybrid space, distinguishable clusters were observable along the objective feature (y-axis in *Figure 5 (c)*).

Combining decision features with the objective space seems to be relevant. Hybrid space is an example. But we can also imagine applying clustering algorithms two (or more) times to re-segment clusters already created, and successively in the objective space and then the decision space, or vice

versa. In this way, we could obtain good clusters from the energy performance point of view and that make sense for the descriptive criteria.

4. CONCLUSION

The results showed the possibility that the clustering techniques might be efficient for the modeling of building stocks considering energy performance of houses. It was verified that, even when the number of clusters was unknown, it could be decided by estimation of clustering performance. Some algorithms turned out to be probably inappropriate in certain data distribution. If noise points exist in the data, less algorithms are capable to handle with them, and the other algorithms considered the noise points as normal data. In the case of extreme noise, which can distort the clustering results, proper preprocessing seems to be required before the clustering analysis.

As a first application of clustering techniques to the building stock data, the results offer the possibility of further studies. Firstly, parametric studies of each clustering algorithm could be considered. For some algorithms, the required parameters are not intuitionally comprehensive nor easily decidable. Secondly, differing from this study, the objective space could be extended to other features such as the entire final energy consumption instead of the final energy consumption for heating, the unique objective feature utilized in this study. Even further, the investment cost and the improvement of energy efficiency might be interesting objective features, though these data require data of energy retrofit scenarios.

5. BIBLIOGRAPHY

Dueck, Delbert. 2009. “Affinity Propagation: Clustering Data by Passing Messages.” University of Toronto.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” *Kdd*.

European Commission. Commission Delegated Regulation (EU) No 244/2012 of 16 January 2012.

Fürnkranz, Johannes, Dragan Gamberger, and Nada Lavrač. 2012. *Foundations of Rule Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-75197-7.

Pearson, Karl. 1904. “Report on Certain Enteric Fever Inoculation Statistics.” *Bmj* 2 (2288). British Medical Journal Publishing Group: 1243–46. doi:10.1136/bmj.2.2288.1243.

Hartigan, John Anthony, and Manchek Anthony Wong. 1979. “Algorithm as 136: a K-Means Clustering Algorithm.” *Applied Statistics* 28 (1): 100. doi:10.2307/2346830.

Murtagh, Fionn. 1983. “A Survey of Recent Advances in Hierarchical Clustering Algorithms.” *Computer Journal* 26 (4). Oxford University Press: 354–59. doi:10.1093/comjnl/26.4.354.

Rousseeuw, Peter J. 1987. “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20 (November): 53–65.

Struyf, Anja, Mia Hubert, and Peter Rousseeuw. 1996. “Clustering in an Object-Oriented Environment.” *Journal of Statistical Software* 1 (4). doi:10.18637/jss.v001.i04.

TABULA Project Team. 2012. “Typology Approach for Building Stock Energy Assessment. Main Results of the TABULA Project.” Edited by Tobias Loga et al. Institut Wohnen und Umwelt.

Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. 1997. “BIRCH: a New Data Clustering Algorithm and Its Applications.” *Data Mining and Knowledge Discovery* 1: 141-182.