Highlights

Developing a Dynamic Carbon Benchmarking Method For Large Building Property Estates

Miles J.S. Gulliford, Richard H. Orlebar, Max H. Bird, Salvador Acha, Nilay Shah

- A novel benchmark method to aid the decarbonisation of buildings is introduced
- Alternative error metrics and a multi-layer model ensure a meaningful benchmark
- Decision Tree regression has lowest average error (1.5 $kgCO_2m^{-2}period^{-1}$) thus highest accuracy
- Decision Tree regression assigns all known poor performers in worst 30% of buildings
- Humidity shown to be second most important variable for natural gas use prediction

Developing a Dynamic Carbon Benchmarking Method For Large Building Property Estates

Miles J.S. Gulliford, Richard H. Orlebar, Max H. Bird, Salvador Acha, Nilay Shah

[a]*Department of Chemical Engineering, Imperial College London, United Kingdom*

Abstract

As supermarkets are known to be very energy intensive, improvements made to their efficiency can enable operators to reduce not only carbon emissions but also costs, in line with corporate and legislative targets. This study presents a novel benchmarking method to appraise emission and cost performances across a portfolio, enabling building managers to identify sites that are underperforming, taking as a case study a large number of food retail stores. Multiple layers, detailed variable selection including weather features and regression technique comparisons (Multivariate Linear Regression (MLR), Artificial Neural Network (ANN) and Decision Tree (DT)), are considered in model construction. Efficiency is evaluated on multiple bases with a focus on emissions. These are clustered together to produce a benchmark to inform investment decision-making across a portfolio. The DT technique was found to be the most effective, producing a benchmark with low average error (1.5 $kgCO_2m^{-2}period^{-1}$) and high maximum error (21 $kgCO_2m^{-2}period^{-1}$) indicating high accuracy and high discernment respectively. This model also correctly classified buildings known to perform poorly into the worst 30% of buildings in the portfolio. This work highlights the need for further research into natural gas consumption benchmarking and particularly the use of humidity data to this end.

Keywords: Artificial Neural Networks, Building Performance, Carbon Benchmarking, K-means Clustering, Multivariate Linear Regression, Decision Tree, Energy Efficiency, Sustainability Strategy, Retrofit Investment, Commercial Buildings

Abbreviation	Full Name
ANN	Artificial neural network
$CBECS$	Commercial Buildings Energy Consumption Survey
CCR	Combined cost residual
CER	Combined emission residual
CDD	Cooling degree day
$CEDA$	Centre for environmental data analysis
CF	Carbon factor
CHP	Combined heat and power
$COVID-19$	Coronavirus disease 2019
DT	Decision tree
HDD	Heating degree day
MAE	Mean absolute error
MLR	Multivariate linear regression
MME	Mean maximum error
MSE	Mean square error
$RFECV$	Cross-validated recursive feature elimination
UK	United Kingdom
US	United States

Symbol	Description	Unit
$CF_{elec,i}$	Electricity Carbon factor of building i	$kgCO_2kWh^{-1}$
$CF_{gas,i}$	Natural gas Carbon factor of building i	$kgCO_2kWh^{-1}$
n	Number of observations	-
$PF_{elec,i}$	Electricity price factor of building i	$£kWh^{-1}$
$PF_{gas,i}$	Natural gas price factor of building	$£kWh^{-1}$
$R_{elec,i}$	Electricity residual of building i	$kWhm^{-2}period^{-1}$
$R_{gas,i}$	Natural gas residual of building i	$kWhm^{-2}period^{-1}$
S_i	Set i of buildings	-
T_i	Average daily temperature on day i	k
T_{bc}	Cooling baseline temperature	k
T_{bh}	Heating baseline temperature	k
u_i	Binary variable indicating if T_i is below T_{bh}	-
v_i	Binary variable indicating if T_i is above T_{bc}	-
$X_{predicted}$	Building predicted energy intensity	$kWhm^{-2}period^{-1}$
X_{real}	Building observed energy intensity	$kWhm^{-2}period^{-1}$
X_i	Variable i for set of buildings X	-
x_{ik}	Output of node k from layer i	-
x_{mn}	MLR variable n for store m	-
$y_{i,predicted}$	Predicted energy use intensity	$kWhm^{-2}period^{-1}$
$y_{i,real}$	Observed energy use intensity	$kWhm^{-2}period^{-1}$
α	Zeroth order coefficient of MLR	-
β_n	First order coefficient of MLR for variable n	-
γ	Zeroth order coefficient of ANN node	-
ϵ_k	First order coefficient of ANN node input from node k	-
μ_k	Centre of K means cluster k	-

1. Introduction

In 2019, building operation made up 30% of global CO_2 emissions with 10 $GtCO_2$ emitted annually [1]. The Intergovernmental Panel on Climate Change estimates that non-domestic buildings could achieve an 18% reduction in carbon emissions through low cost efficiency improvements [2], they cite this "untapped energy efficiency potential" as one of the most important pathways to reduce building emissions [1]. The food retail industry is large and complex providing the vital service of supplying goods to society. Supermarkets contribute to approximately 1% of the UK's GHG annual emmissions [3], in 2020 this was equivalent to 4.1 $MtCO_2e$ [4]. The estimated revenue of the food industry in 2020 was £12.8 billion [5]. In particular, food retail buildings are energy intensive due to features including, lighting, heating, ventilation, air conditioning and refrigeration required to maintain customer comfort and food freshness [3]. Other features including large heating volumes, high operational intensity and outdated insulation have been identified as contributors to increased intensity of supermarket operation [6]. Supermarkets are therefore a building class that require careful monitoring to ensure resources are used efficiently [7]. Due to their high energy use and strong economic presence this paper focuses on analysing the performance of food retail buildings.

To identify the best set of solutions to reduce emissions and costs, it is necessary to understand the complex load requirements of power, heating, and cooling of energy intensive buildings [8]. Through careful analysis and deep understanding of energy demand, organisations can understand the impact of their investments and begin to devise a low emissions roadmap to meet long term environmental targets and introduce technical solutions [9]. Hence, it is evident that modelling and benchmarking tools for building performance can inform decision-makers on best carbon mitigation strategies.

There exists a significant amount of literature on the analysis of building energy use and performance efficiency. Technical, statistical and data-driven techniques have been developed for energy demand prediction [10]. A good example of this at the city scale is the work of Larivière and Lafrance, who present a model of urban energy consumption based upon city characteristics, to aid future planning and design for efficient cities [11]. Other works include tools for diagnosis and retrofitting of individual buildings [12, 13]. Here Kalogirou and Bijou's work using neural networks to evaluate the implementation of solar energy to new designs stands out for its clear focus on a specific, singular application [14]. Furthermore, extensive work exists in the field of building performance benchmarking [15, 6]. Within the building sector, benchmarking refers to the comparison of energy performance across similar buildings that share common attributes [16]. These typically result in a performance score or rating. Benchmarking is an essential and distinct component of the building performance modelling field as it emphasises building comparison over model accuracy and hence is better aligned with industry use. While a highly accurate energy use intensity model may be useful for forecasting investment scenarios, it is not calibrated to inform end-users how efficiently a building is performing against its cohort. Hence, in industrially focused works such as this, benchmarking is typically the researcher's tool of choice. A number of white-box methods [17] have been developed based on detailed industry knowledge. The most renowned of these models being the Energy Star methodology [18], a benchmark designed by the US Environmental Protection Agency, whereby each building is assigned a 1-100 rating for energy performance with the goal of justifying building upgrades and improving performance. While white-box initiatives such as these act as a useful form of performance disclosure, their reliance on detailed technical information means they are less suited to evaluating building performance at scale. Therefore, by leveraging improved reporting across building portfolios, black-box data-driven methods offer significant potential for large building datasets [19]. Although such methods are in increasing demand, due to their ability to use generic information to identify meaningful insights, such methods have rarely been developed and demonstrated for commercial building portfolios.

Data-driven methods employed for building energy performance prediction and benchmarking include multivariate linear regression (MLR) [15], decision trees (DT) [20], and K-means clustering [6]. More nuanced and novel approaches include work by Lara et al. [21] where K-means clustering was used to distribute 59 schools across Italy into three groups by a set of attributes, before energy use prediction through MLR is applied. This resulted in an improved prediction with a higher coefficient of determination (R^2). Yalcintas' 2006 work used aritifical neural networks (ANNs) to benchmark and predict energy demand from lighting, plug loads and HVAC systems in over 60 Hawaiian school buildings, while modelling how retrofitting efforts would affect these benchmarks [22].

Gao and Malkawi's K-means clustering based benchmark, evaluated relative importance of variables for electricity use prediction, utilising 2,000 buildings from the US' commercial buildings and energy consumption survey (CBECS) [6]. This work also incorporated heating and cooling degree day (HDD/CDD) metrics designed to quantify the sensitivity of heating and cooling loads over a given time period. However, this analysis indicated these weather variables showed limited predictive value. Similar findings were observed by Spyrou et al. in their regression based energy prediction models, this work is also noted for being one of few using data from a commercial food retailer [23]. It is somewhat surprising that weather data have historically shown little value to energy benchmarking methods covering long-periods of time and diverse climates. Therefore, previous works have suggested that effective implementation of weather features may add value to the academic literature [19].

Several works in the literature have contrasted and evaluated different regression techniques for energy demand prediction within benchmarking applications. Ding and Liu for example, compared the Energy Star method, China's national energy consumption standard and stochastic frontier analysis [24]. Here, benchmark value was evaluated in terms of consistency across the models, and the differences in categorisation were considered carefully. It should be noted that the selection of models used focused on applications for policy makers, as opposed to building managers. Meanwhile, Tso and Yao compared MLR, ANN and DT regression models for building energy consumption prediction [25]. This was considered to add significant value to the field, for future prediction methods. However, model construction and performance evaluation were undertaken with emphasis upon accuracy, rather than the strategic value of the model as a diagnostic tool.

This work builds upon the wealth of existing benchmarking research to develop a novel, industrially applicable method for the evaluation and identification of emissions reduction opportunities across a building portfolio using data provided by a major UK retailer. Several opportunities were identified to produce an enhanced, and hence more industrially relevant benchmark. First, by providing efficiency scores on both carbon and cost bases, greater insight is provided decision makers, leading to more efficient and rapid decarbonisation. Efficiency benchmarks in financial terms have not been produced in this field previously, and localised energy cost data used in this work ensured this output could be produced at high quality.

Additionally, the implementation multi-layered regression model similar to the work of Lara et al. [21], for benchmarking outcomes was identified as an opportunity to produce a more accurate model providing more actionable insight than previous benchmarks. To further the work of Lara et al. an outlier removal layer within the regression is proposed to reduce model skew. Finally, particular attention is paid to the enhanced implementation of weather variables such as humidity, a previously unused variable. This capability enhanced the benchmark's ability to assess building performance, accounting for seasonal or climate factors alongside variables designated as strategically unmanageable, including building age and size. Further, this work evaluates the performance of energy prediction regression methods specifically for commercial benchmarking applications. This has been highlighted in literature as an area for further work [20, 26]. Three regression techniques for a commercial benchmarking approach were assessed; MLR, ANN and DT. This comparison offered compelling insights into relative variable importance, and the inherent differences between these techniques, ensuring an incisive final model.

This paper is composed of five sections. The current section has provided the motivation, background, purpose, and scope of the problem. The second section describes the methodology, mathematical formulation of the model and the means through which it was evaluated. The third section provides the results from the electricity and natural gas consumption benchmarking. It goes on to discuss the differences in results and variable importances derived from the different regression techniques. The fourth section discusses the trade-offs between different regression techniques and the limitations of the benchmark. The last section of this work provides concluding remarks.

2. Methodology

The benchmark method was developed through several steps. First, careful data collection and processing was undertaken. Next, a variety of data driven techniques were implemented in a multi-layered structure to produce the final benchmark. Finally, the performance of different benchmark were compared. The methods behind each step are elucidated below.

2.1. Data Collection and Treatment

2.1.1. Building Data

A list of buildings and their attributes was sourced from a commercial food-retailer. Their respective electricity and natural gas consumption in four-week periods, spanning financial years beginning 2017 through 2019, was collected and appended. This dataset contained details of approximately 2,000 buildings. It was desired to study buildings of a single category type, as this ensured similar assets and operations were compared. As large food retail sites were identified as having higher energy requirements [23], only properties categorised by the business as supermarkets were considered. This left 584 buildings available for the undertaking of this study.

2.1.2. Weather Variable Creation

Hourly temperature and humidity data from across the UK were sourced from the Centre for Environmental Data Analysis (CEDA) [27]. These datasets were loaded into an SQLite database from which they could be called. Data was collected for each building by matching its postcode location to the closest weather station.

Weather variables in CEDA include: hourly and daily average temperature, and humidity. Daily temperature data was converted to heating degree days (HDD) and cooling degree days (CDD), as shown in equations 1 and 2 respectively. The basis of these degree day metrics is to measure the cumulative days when temperature is above or below a given baseline, and the magnitude of deviation from this baseline. These may be interpreted as demand on heating and cooling systems over the time period considered, for example HDD is expected to be positive in the winter, and zero in the summer. These values are commonly used in literature [22] as they reflect more accurately weather effects on energy demand.

$$HDD = \sum_{i=day} v_i(T_i - T_{bh}) \quad v_i \begin{cases} 1 & T_i < T_{bh} \\ 0 & \text{Else} \end{cases} \tag{1}$$

$$CDD = \sum_{i=day} u_i(T_i - T_{bc}) \quad u_i \begin{cases} 1 & T_i > T_{bc} \\ 0 & \text{Else} \end{cases} \tag{2}$$

To maximise the value of the HDD and CDD datasets, analysis of their correlation with electricity was completed for different baseline temperatures T_b across a range of 10-20 °C. Correlation was measured as the R^2 of linear regression of HDD and CDD with electricity consumption. Baseline values of 15.5°C for CDD and 12°C for HDD were selected. These were validated as appropriate baseline values through consultation with industry experts on standard baselines for supermarket performance analysis specific to the UK.

2.1.3. Carbon and Price Factors

A carbon emissions factor (CF) represents the CO_2 produced per unit energy consumed ($kgCO_2kWh^{-1}$). A specific CF provided by the industrial partner was used to convert the energy consumption in kWh of each building to a carbon emissions basis. As CF values were only provided for the financial year 2017, the electricity CF values were reduced at a constant rate of 10% per year, based on the rate of decarbonisation of the UK's energy grid in 2019 [28]. It is noted that this forecasting approach reduced model carbon accuracy, this is designated an opportunity for model improvement. On the other hand, a constant natural gas CF of 0.184 $kgCO_2kWh^{-1}$ was taken from government data, as negligible change was observed across the period [28]. Similarly, electricity and natural gas price factors (PF) with units of $£kWh^{-1}$ were implemented to convert energy consumption to an economic KPI. These provided tangible metrics of carbon and cost intensity for each building ($kgCO_2m^{-2}period^{-1}$ and $£m^{-2}period^{-1}$). It should be noted CCR and CER values are given on a per-period basis, where a period is a 28 day interval, corresponding to the resolution of consumption data used in this study.

2.1.4. Variable Selection

Once all datasets were collected, qualitative and quantitative methods were used to eliminate variables which would not meaningfully contribute to model performance. A full list of variables available is provided in appendix Appendix A.1.

Factors considered easily improvable, such as light fitting type, refrigerant type or CHP capacity, are frequently excluded from energy intensity benchmarking models in literature [7]. This is to avoid penalising improved buildings. This reasoning is best understood through the consideration of two identical buildings, one with efficiency improvements made to its manageable variables, and one without. In the case that these variables are not considered in the energy use prediction model, both buildings will be predicted to perform at the same level. Hence, the unimproved building will be identified as underperforming against predictions and thus will be assigned higher priority for improvement. In the case that the model accounted for improvable variables, it is possible the improved building would be assigned higher priority for investment, despite having a smaller margin for improvement. Therefore, factors considered as easily improvable or attainable were excluded from the prediction model in this work.

Next, variables with low variance (below 5%) were removed from consideration, as they reflected variables that changed little across the entire database, providing limited predictive insight. Likewise, variables with high degrees of co-variance were removed. This prevented the amplification of a single characteristic that would otherwise be captured in another variable: cafés and petrol stations were removed here as they were found to be strongly ($r > 0.6$) associated with building sales area.

Analytical variable selection was undertaken using a cross-validated recursive variable elimination (RFECV) algorithm for the MLR model regression. Variables were scored on their importance to the model (see section 2.4.3) by assessing its coefficient from an MLR regression (see section 2.2.2), the lowest value variable was removed and the process was repeated. The number and combination of variables, which minimised model error (Mean absolute error, MAE, see equation 10) was selected for the MLR regression. The variables removed in each step and final variables used in each model can be found in Appendix A.1.

2.2. Data Science Techniques

Data-driven techniques were employed in this work to provide insight unavailable through traditional methods. The techniques used are elucidated below.

2.2.1. K-Means Clustering

K-means clustering is a popular method for grouping m items into k clusters in which each item belongs to the cluster with the closest centroid. Given a set of observations $(x_1, ..., x_m)$, where each contains n variables, K-Means clustering defines sets $S = S_1, ..., S_k$ with d-dimensional centroids μ so as to minimise the sum of square error within the group (i.e. variance):

$$arg_S min \sum_k \sum_{\mathbf{n}_k} || \mathbf{x} - \mu_k ||^2 \tag{3}$$

2.2.2. Multi-variate Linear Regression (MLR)

MLR closely fits an $n-1$ dimensional hyperplane for m observations in an n dimensional space. The n dimensions correspond to variables evaluated, where the n^{th} attribute is predicted by the $n-1$ dimensional hyperplane. This may be expressed as an equation of the form shown in equation 4 below.

For example, a two-dimensional plane may be defined in a three-dimensional space, with the first two dimensions corresponding to attributes area and age. Here, the plane would specify, or predict the third dimension, electricity consumption.

$$y_i = \alpha + \sum_{n=1}^{n-1} \beta_n x_{mn} \tag{4}$$

The coefficient of each variable β_n and intercept α are determined to minimise a cost function equivalent to model error, in this case mean squared error (MSE), as given in equation 5.

$$MSE = \sum_{m=1} (y_{i,predicted} - y_{i,real})^2 \tag{5}$$

2.2.3. Artificial Neural Network (ANN)

Similarly to the MLR method, the ANN regression predicts values of an n^{th} variable from $n-1$ variables after fitting with m observations. This is achieved by passing the values of the $n-1$ attributes through a network of nodes. The output of this network is the predicted value of the n^{th} variable. This structure is based on the way the brain learns [29].

There are three types of nodes in an ANN: input nodes, hidden nodes and output nodes. Each input node, sits in the first layer, and takes the value of a variable for a given observation. Each output node sits in the final layer and yields a predicted value.

These two groups of nodes are connected by a network of j layers of hidden nodes. In this work, the hidden nodes are arranged in a feed-forward structure. Here, nodes are arranged such that each node in layer i takes inputs from all nodes in layer $i-1$ and transform them according to a series of weights, as shown in equation 6.

$$x_{i,j} = \gamma + \sum_{k=1} \epsilon_k x_{i-1,k} \tag{6}$$

Where $x_{i-1,k}$ is the output of node k from the previous layer, and $x_{i,j}$ is the output of the node. γ and ϵ_k are weights assigned to minimise the model's cost function, in this work MAE was used as shown in eq. 10, as squared error metrics amplify the influence of outliers.

2.2.4. Decision Tree (DT)

Decision trees, when applied to continuous outcomes, are referred to as regression trees. These work by partitioning data into subsets via a series of binary splits using Boolean logic. Each point at which a decision is made is referred to as a node.

The Classification and Regression Tree algorithm was selected for decision tree construction, as it accepts continuous inputs and outputs, and has been employed for benchmarking outcomes previously [20]. For regression models, each split was made to minimise error (MAE) between the mean of the generated subset and each of its datapoints. MAE was chosen to minimise the influence of outliers as discussed in section 2.2.3 above.

2.3. Building Energy Performance Benchmark Modelling

2.3.1. Model Structure

The benchmark model has been developed based on a combination of previous works [6, 24, 21], whereby electricity and natural gas consumption is predicted for each building using regression techniques. These predicted values are designed to be specific efficiencies at which each building could operate. The difference between the actual and predicted consumption values are labelled consumption "residuals", these are obtained as shown in eq. 7. These are used to indicate the performance level of a building with respect to its predicted consumption, where greater than predicted energy use yields a positive residual. Positive residuals indicate opportunities for improvement in efficiency. Additional steps included in this model are: outlier removal, K-means clustering prior to regression for MLR and ANN models and final priority clustering by investment KPIs. A detailed step-by-step outline of the model structure is shown in figure 1.

$$R = X_{Real} - X_{Predicted} \tag{7}$$

While the three models used different regression techniques, to maintain valid comparisons a consistent model structure and output was defined. This was made up of four steps as detailed below and in figure 1. K-means clustering, was used in steps one and three, to distribute supermarkets into manageable groups with similar variables prior to regression modelling to improve model resolution. This is based on work by Lara et al. [21]. As the decision tree model created its own subsets, K-means clustering prior to implementation was deemed unnecessary in this case. Regressions were undertaken on a period-by-period basis to avoid repeat observations of the same buildings which were found to lead to overfitting. Buildings which did not use gas were not included in the natural gas consumption regression.

1. **Data Pre-processing:**
   Prepare the data to be fed into the model. This included the collation of weather data and building characteristics.
2. **Removal of Outliers on Carbon Residual basis:**
   Reduces the impact of significant outliers on the regression model.
   (a) K-Means Clustering:
   Clustered the buildings into 3 groups based on characteristics such as size or age.
   (b) MLR Fitting & Prediction Modelling:
   Generates initial benchmark values for the entire dataset using an MLR method.
   (c) Carbon Residuals Generated:
   The carbon residuals were calculated as the difference between the actual consumption values and a buildings benchmark.
   (d) Buildings with residuals outside of 2 standard deviations are removed:
   Removes buildings with large acting residuals. Passes 95% of buildings to step 3 for training.
3. **Main Benchmark Iteration:**
   Final residual metrics generated for use in step 4 and analysis.
   (a) K-Means Clustering:
   Models One & Two: Clustered the buildings into 3 groups based on characteristics such as size or age.
   Model Three: Step skipped as Decision Tree performs clustering step inherently.
   (b) Fitting & Prediction Modelling:
   Different regression techniques were used in each model to generate benchmark values for each building:
   Model One: Multivariate Regression
   Model Two: Artificial Neural Network
   Model Three: Decision Tree
   (c) All Residuals calculated and stored:
   Residuals were calculated as the deviation from the predicted metrics such as electricity consumption or carbon emissions.
4. **Priority K-Means Clustering:**
   Final investment priority groupings were developed with respect to investment criteria such as cost or carbon residuals.

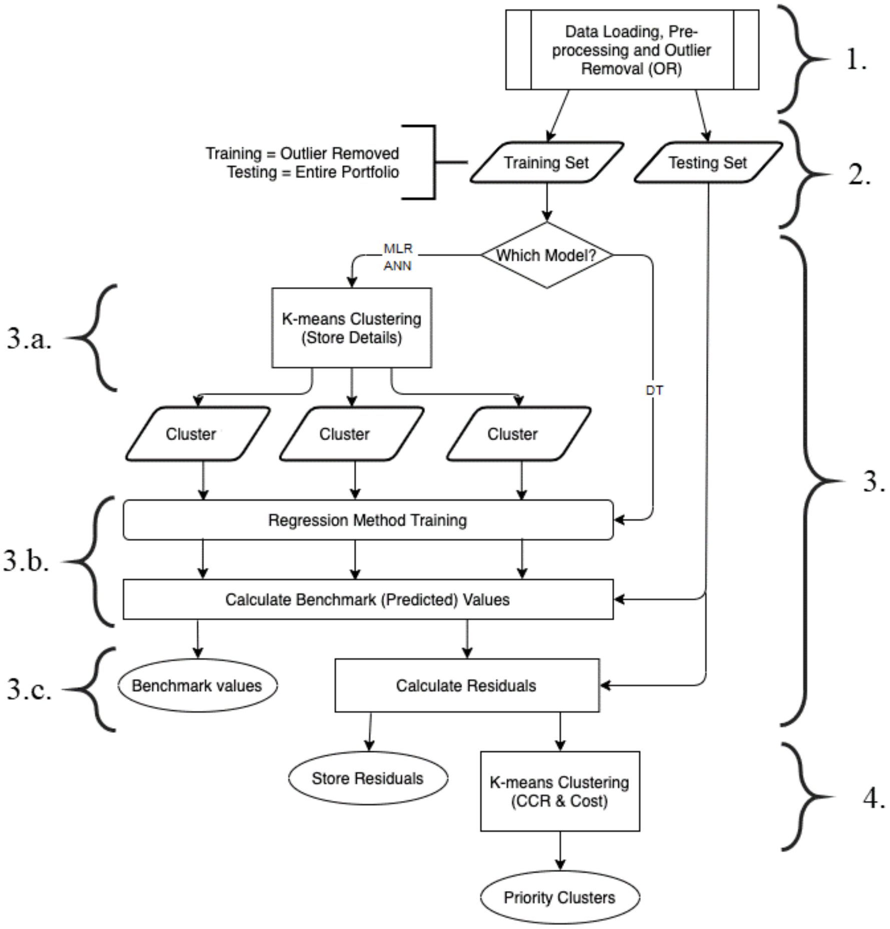
Data Loading, Pre-processing and Outlier Removal (OR)
1.
Training = Outlier Removed
Testing = Entire Portfolio
Training Set
Testing Set
2.
Which Model?
MLR
ANN
DT
3.a.
K-means Clustering (Store Details)
Cluster
Cluster
Cluster
3.
3.b.
Regression Method Training
Calculate Benchmark (Predicted) Values
3.c.
Benchmark values
Calculate Residuals
Store Residuals
K-means Clustering (CCR & Cost)
4.
Priority Clusters

Figure 1: Model Structure Flowsheet

As all model construction algorithms (see sections 2.2.2, 2.2.3, 2.2.4) aim to minimise prediction error, buildings with abnormal electricity or natural gas demands have greater influence over the regression models. This was considered undesirable as significant deviations would lead to unrealistic benchmark values for all buildings. Therefore, in step two, buildings beyond two standard deviations from the mean were removed, leaving approximately 95% of buildings as the training set. This was to limit the influence of outliers on the predicted carbon residuals.

2.3.2. ANN Model Specifications

Where ANN regression is used, a network with one hidden layer and nodes equivalent to half the number of input nodes was selected as appropriate based upon previous ANN models developed for benchmarking purposes [26, 22]. In this case the separate electricity regressor had 16 input nodes and the natural gas regressor had 15. Both had a single output node, corresponding to the predicted performance value. The feed forward design of ANNs in this work present no memory or competition between nodes.

2.3.3. DT Model Specifications

Decision trees use short-sighted decision-making processes, meaning each split was optimised without consideration of decisions further down the tree. A maximum of five splits and minimum leaf size of 5% of buildings were specified to avoid overfitting. These values were specified based on previous building energy efficiency benchmarking works [20, 30].

2.3.4. Model Output

Within each model, a period-specific and unique benchmark value for electricity and natural gas demand was calculated as the deviation of the building's consumption from its predicted value.

Combined Emissions Residual (CER) and Combined Cost Residual (CCR) were identified as powerful KPIs to compare buildings in terms of their environmental and economic impact respectively. These metrics were inclusive of natural gas and electricity components. CER were calculated as a sum of carbon residuals (the products of each consumption residual and their respective carbon factors), accounting for both electricity and natural gas over the entire period considered as shown in equation 8. Likewise, CCR were calculated using price factors (PF) for electricity and natural gas, as shown in equation 9.

$$CER = \sum_{Time}^{i} (\mathrm{R}_{Elec,i} \times CF_{Elec,i} + \mathrm{R}_{Gas,i} \times CF_{Gas,i}) \tag{8}$$

$$CCR = \sum_{Time}^{i} (\mathrm{R}_{Elec,i} \times PF_{Elec,i} + \mathrm{R}_{Gas,i} \times PF_{Gas,i}) \tag{9}$$

When plotting the outputs of each model, minor differences were difficult to observe visually due to the large number of datapoints included. However, it was clear that for each model there existed buildings which lay outside of the high density regions; these were successfully identified by the priority clustering step which highlighted similar numbers of extreme poor-performers across the models, as shown in priority cluster one in table 1. Similar clustering patterns emerged for each model. These were arranged from worst to best by euclidian distance of the centroids with respect to CCR and CER from the best performing observation. Using these priority clusters to identify smaller groups of stores, the model succinctly indicates where the performance of a building falls within its portfolio with respect to costs and emissions.

2.4. Model Testing

To create a rounded view of each models applicability, a variety of evaluation methods were used. This section explains how each model performed.

2.4.1. Model Error

Model error is a common means of understanding and comparing model performance [26, 30]. Error metrics were taken as averages for each model. As MAE was utilised in construction of the ANN and DT regressions, it was deemed the most relevant metric for model evaluation. This represented the model's ability to realistically predict energy consumption. However, as residuals, equivalent to model error, are a key model output, complete minimisation was not desired, this is explored further in Results section 3.2.1. MAE is defined by equation 10 as:

$$MAE = \frac{\sum_i |y_{i,predicted} - y_{i,real}|}{n} \tag{10}$$

$$MME = max(|y_{i,predicted} - y_{i,real}|) \tag{11}$$

To supplement MAE, mean maximum error (MME), given in eq. 11, was collected for each model to provide insight into the maximum deviation from a benchmark value; this indicated the harshness of the benchmarks.

2.4.2. Model Validation

A list of buildings with known under-performing supermarkets was identified and collated by the commercial partner. The benchmarks of each of these buildings were qualitatively analysed to ascertain the realism and validity of each model, by considering if such buildings were in line with previous works. A full description of each building can be found in the Appendix section A.6.

2.4.3. Variable Importance

The relative importance of the variables used in each model were determined. The permutation importance method was employed, as it can be applied to all regression methods ensuring comparable outcomes. [31]. This method determined the importance of variable n by measuring the impact on model error (here MAE), when observations of variable n are randomly shuffled between buildings.

3. Results

In order to understand the effectiveness of the benchmarking methodology developed above, the different models were evaluated against one another. Additionally, the benchmarks allocated to case-study buildings were considered as a means of model validation. Finally, the importance assigned by the models to different variables offered further insight into differences between model prediction methods.

3.1. Model Outputs

Table 1 shows the distribution of buildings between clusters for each model. Although the models showed similar distributions of buildings between clusters, it was observed that the DT model distributed buildings somewhat less evenly between clusters with a variance of 3.9%. Further, the DT model provided a more discerning prediction of the portfolio, with the largest proportion of buildings (52.7%) categorised as under-performing. Meanwhile, the MLR model was the most benign or optimistic with 49% of buildings categorised as over performing, and showed more evenly distributed clusters with a variance of 1.7%. This was due to the inherent differences between regression methods. While the MLR method was constructed using MSE, which emphasised anomalies during model fitting, the

ANN and DT methods minimised MAE, which weighted anomalies less, resulting in higher residuals for these buildings. Hence, the use of MAE did ensure a more discerning model, a better fit to inform and guide decision-makers.

Table 1: Distribution of buildings by priority cluster and over vs under performance by model

	Number of buildings			**Proportion of buildings %**		
Priority Cluster	**MLR**	**ANN**	**DT**	**MLR**	**ANN**	**DT**
1	6	5	3	1.0	0.9	0.5
2	97	40	13	16.7	6.9	2.2
3	190	181	175	32.6	31.1	30.1
4	186	259	277	32.0	44.5	47.6
5	103	97	114	17.7	16.7	19.6
Performance						
Over	285	279	275	49.0	47.0	47.3
Under	297	303	307	51.0	52.1	52.7

3.2. Model Comparisons

3.2.1. Model Error

The average MAE and MME of each model for both natural gas and electricity were computed, as shown in table 2.

The DT regressions yielded the lowest MAE, averaging 1.49 $kgCO_2m^{-2}period^{-1}$, followed by the ANN model for electricity, while the two showed similar MAE values for natural gas regressions at around 1.7 $kgCO_2m^{-2}period^{-1}$. The MLR model consistently scored the highest MAE, with an average of 1.68 $kgCO_2m^{-2}period^{-1}$. It should be noted that the *period* unit corresponds to 28 days, as discussed in section 2.1.3. The differences in error can be partially explained by the fact that the error MLR minimised was MSE, hence it was expected to achieve a sub-optimal MAE. However, the

Table 2: MAE and MME for overall models, electricity regressions and natural gas regressions

Model	MAE $kgCO_2m^{-2}period^{-1}$	MME $kgCO_2m^{-2}period^{-1}$
MLR	1.68	14.9
Electricity	1.50	11.5
Gas	1.86	18.3
ANN	1.53	15.0
Electricity	1.38	11.3
Gas	1.67	18.6
DT	1.49	21.1
Electricity	1.27	13.6
Gas	1.70	28.6

discrepancy also suggested that the relationship between the variables and electricity and natural gas consumption intensity showed a non-linear behaviour, which the ANN and DT models were able to capture, unlike the MLR [24]. These relationships are similar to the results and conclusions drawn by Yalcintas and Ozturk, in their work predicting energy consumption for buildings from the CBECS using ANNs [26].

MAE was consistently higher for natural gas regressions than for electricity. One explanation may be that natural gas demand was less comprehensively explained by variables considered here. For example, Mavromatidis et al. [12] research on energy use in a supermarket by service, uses hourly weather and indoor building average temperatures as inputs for predicting boiler energy usage. However, such data was not available during the development of these models.

MME gave a strong indication of the maximum size of residuals produced by each model. Despite producing the lowest MAE, the DT model showed the highest MME, an average of 21.1 $kgCO_2m^{-2}period^{-1}$. This was expected, as the DT model assigned a mean benchmark value to all buildings within a given leaf. As these values were less tuned to each individual building, greater errors were produced for anomalies. This approach seemed to produce greater insight as buildings with the greatest need for investment were assigned more distinct residual values. This also explained the less even distribution of buildings between clusters in the DT model, see table 1.

3.2.2. Model Validation

While co-comparison of models offered insight into their relative performances, it was also necessary to validate benchmarks using case study buildings. Therefore, the ability of the models to diagnose the poorly performing buildings (A-D) from Appendix Section Appendix A.2 was taken as a qualitative measure of each model's validity. The CER, emissions percentile and benchmark score for each case-study building are given in table 3.

Table 3: Carbon residual, percentile rank with lowest percentile as worst and priority cluster rank of sample buildings

	Carbon $_{kgCO_2m^{-2}period^{-1}}$			**Percentile %**			**Cluster Rank**		
Building	**MLR**	**ANN**	**DT**	**MLR**	**ANN**	**DT**	**MLR**	**ANN**	**DT**
A	30.2	29.9	149	34.4	36.4	4.8	5	3	4
B	-23.9	-7.73	107	60.8	57.4	9.3	4	3	3
C	58.2	21.8	61.5	23.5	40.0	21.5	3	4	3
D	-44.6	-47.8	45.9	68.2	76.1	26.6	4	4	3

Buildings A-D were manually identified by the commercial partner as amongst the worst performing. While buildings which were known to be leaking (A & C, see Appendix section Appendix A.2) were registered as poor-performing by all models, only the DT model identified all case study buildings as under-performing with respect to the carbon residuals.

A point of concern was that models MLR and ANN suggested that buildings B & D were actually over-performing, given building B was identified for high demand and D for age. It was found that these models predicted higher than observed natural gas consumption, resulting in favourable (i.e. negative) gas residuals. This effect dominated CER values as natural gas is more carbon intensive. This highlighted the value of the DT model for decision-making as its harshness meant that electricity consumption above predicted levels was still highlighted despite lower-than-predicted natural gas use.

Furthermore, the DT model ranked these buildings within the worst 30%. Hence, the DT model was considered the most valid, and therefore useful for decision-makers, as it most appropriately categorised known worst performers. Additional discussion of case study buildings may be found in appendix section Appendix A.2.

3.2.3. Variable Importance

The relative importances of variables for natural gas and electricity prediction regression are given in figures 3 and 2.

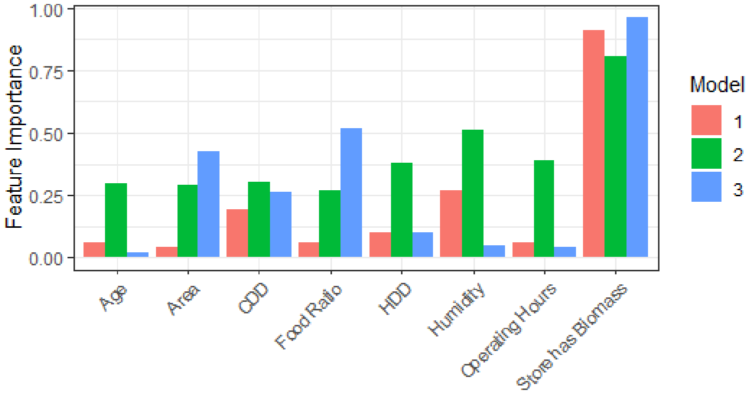

Figure 2: Variable importance by model for natural gas use regression

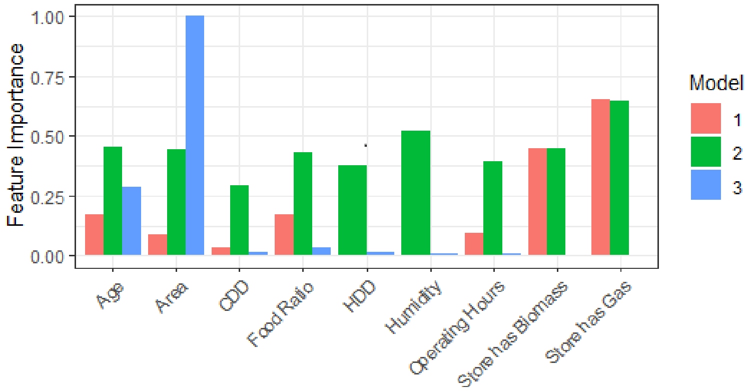

Figure 3: Variable importance by model for electricity use regression

As the models were applied across an extended operation period (3 years), weather variables were of particular interest to minimise seasonal effects upon the model.

Humidity is a variable yet to be implemented in building performance benchmarking. While humidity had been identified as a high value energy use predictor [32], figure 2 provided further support for this observation as it was identified as a variable of second highest importance for MLR and ANN models for natural gas use prediction. Humidity was concluded to be a candidate for further investigation due to its complex relationship with experienced temperature and heating efficiency.

As heating was assumed to be the primary use of natural gas within supermarkets, weather variables were expected to be of greater importance to gas regression models than electricity. This assumption appeared valid for the MLR and DT models, where weather variables were of negligible importance for electricity regression. On the other hand, the ANN model assigned moderate to high importance to all three climate variables. As the other models and previous work [23, 6] suggested weather variables should be of limited value for electricity use prediction, the fact that the ANN model derives value from these attributes suggested it was overfit rather than deriving meaningful insight. However, ANN models are more able to capture non-linear relationships in data. It was therefore considered possible that those variables less valued by other models may simply have had a non-linear relationship with energy use intensity. Hence, further investigation was required.

A number of points suggested that the ANN model was indeed overfit. First, despite appearing to have extracted insights from variables unused by the MLR and DT models, the ANN model error was of a similar magnitude to the other models, as shown previously in table 2. Furthermore, an investigation of variable importance across different time intervals indicated that this model was assigning importance to variables less consistently, as shown in Appendix Section Appendix A.3. This implied the data was not being interpreted in a physically meaningful way, rather the ANN model was forcing value from these variables. Finally, the low volume of data per regression, with the 582 buildings separated into three clusters before regression, was considered a potential cause of overfitting. It was therefore concluded that a degree of overfitting was at play in the ANN model.

The binary variable indicating if a supermarket had a biomass furnace was extremely important to all models for natural gas regression. This was unsurprising as biomass boilers reduce demand on gas boilers. It also offered some explanation as to the poor performance of natural gas regression observed in section 3.2.1, as the models appear to over-rely on a low resolution variable which cannot produce accurate predictions.

A key difference in variable importance was observed in figure 3. The MLR and ANN models selected the binary indicator for natural gas use as their most important variable for electricity use prediction, while the DT model selected area. This was attributed to two factors. First, the model optimisation algorithm for the DT model is distinct in that it makes short-sighted decisions for each branch, while MLR and ANN weigh each variable at once. Additionally, it was considered possible that the relationships between area, age and electricity usage were multimodal rather than continuous, hence the splitting of buildings by these variables as in the DT model would be of greater value than the assignment of a coefficient.

The low MAE, high MME and correct identification of buildings with known performance by the DT model was understood to be more impressive given the small number of variables used to achieve these results. This was considered more applicable as only a small number of variables were required to produce a robust model. Furthermore, the consistency of variable importances suggested physically meaningful interpretation of data was achieved by the DT, with an average variance of 0.05% for electricity regression. Variable importance variances are given in full in Appendix A.3.

Finally, as variable importance did not seem to vary by model error metric (MSE for MLR, MAE for ANN and DT), it was concluded that the implementation of MAE had not impacted model interpretation of variables. Hence MAE was concluded to produce a more discriminating and hence interpretable model, with minimal impact on physical meaning extraction.

4. Discussion

The strategic variable selection approach, and layered model structure employed here, focused the energy prediction benchmark on physically achievable efficiency improvements to aid building managers in allocation of decarbonisation investments.

With the comparison of regression techniques having shone light on their respective strengths and weaknesses, it was concluded that the DT based model performed most effectively. Furthermore, its more scrupulous nature suits it well as a tool for clear decision-making. The ease of implementation and interpretation of this technique, shown both here and in other literature [20, 30] go further to highlight the relevance of DT for benchmarking outcomes.

The introduction of K-means clustering of investment KPIs, illustrates a holistic approach to benchmarking for sustainable investment. While emissions and costs are considered here, further KPIs may be introduced. A manager of a portoflio of buildings could use this tool both to identify targets for decarbonisation investments and study broader trends within their portfolio.

While the model was validated, a key shortfall was poor performance of natural gas use prediction. This is of particular concern as natural gas will tend to dominate emissions as the UK's electricity grid decarbonises. There is therefore potential for further work in gas use prediction.

This benchmark was developed for a single commercial portfolio. The possibility must be considered that further tuning, or an entirely different approach may be required for a portfolio of different size, or nature.

Finally, while this tool has value for short term decision-making, forecasting is necessary to enable the development long term decarbonisation strategies. A forecasting capability would therefore add significant value to this method.

5. Conclusions

An innovative, multi-stage benchmarking method was developed using energy demand prediction models. The implementation of building-specific carbon factors ensured that emissions, rather than simply energy usage, were minimised when producing unique benchmark values. This allowed for the provision of quantitative, realistic and comparable carbon and cost figures. A key unique feature of the benchmarks was the insightful priority clustering of model residuals. This enables decision-makers to rapidly identify groups of buildings requiring improvement with respect to investment KPIs. This clustered output approach was found to correctly identify known poor performers as within the worst 30% of supermarkets in the portfolio for the selected DT model. Furthermore the model produced maintained high accuracy with only 1.49 $kgCO_2m^{-2}period^{-1}$ MAE, but clearly highlighted poor performers with an MME of 21.1 $kgCO_2m^{-2}period^{-1}$. A strategic, layered approach to model design ensured the data-driven methods employed in this work translated into practical insights. This included distinction between improvable and non-improvable variables and emphasis on high carbon standards through rigorous outlier removal and use of anomaly-minimising error metrics. This resulted in a more meaningful model able to rapidly highlight investment targets to building managers.

This work supports previous conclusions on the implementation of weather data for benchmarking outcomes. The analysis and selection of baseline temperature for degree day metrics offered an opportunity for creating energy demand prediction models more robust to disparate climates. Humidity was identified as a particularly important variable for natural gas intensity prediction which had yet to be implemented in other benchmarking applications. Opportunity for further investigation into humidity as an energy use predictor was highlighted.

Another area identified as a shortcoming of this research, and hence a target for further work was each regression's notable limitations in natural gas demand prediction. Variables with potential to improve gas regression include humidity, as well as hourly weather data and average indoor temperatures [12]. The implementation of decreasing electricity carbon factors in line with the decarbonisation of the grid highlighted the increasing relevance natural gas energy intensity will experience. Additionally, while a number of regression techniques were investigated in this work, a comprehensive optimisation of their hyperparamaters was not undertaken. It is suggested, particularly in the case of the favoured DT model, that more rigorous model tuning may add significant value to a benchmarking model similar to this.

References

[1] Thibaut Abergel and Chiara Delmastro, Tracking Buildings 2020 – Analysis - IEA (2021).
URL `https://www.iea.org /reports/tracking-buildings-2020`

[2] C. D., Energy efficiency in buildings, Tech. Rep. ISBN 978-1-906846-22-0, The Chartered Institution of Building Services Engineers, London, SW12 9BS (2012).

[3] M. Hart, W. Austin, S. Acha, N. Le Brun, C. N. Markides, N. Shah, A roadmap investment strategy to reduce carbon intensive refrigerants in the food retail industry, Journal of Cleaner Production 275 (2020) 123039. doi:https://doi.org/10.1016/j.jclepro.2020.123039.
URL `https://www.sciencedirect.com/science/article/pii/S0959652620330845`

[4] G. Smalldridge, 2020 uk greenhouse gas emissions, provisional figures, National Statistics, OGL, London, UK, 2021, p. 1.
URL https://assets.publishing.service.gov.uk/government/uploads/system/upload

[5] Retail: Food Beverages in the UK 2020 — Statista (2020).
URL https://www.statista.com/study/42061/retail-food-and-beverages-in-the-uk/

[6] X. Gao, A. Malkawi, A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm, Energy and Buildings 84 (2014) 607–616. doi:10.1016/J.ENBUILD.2014.08.030.
URL https://www.sciencedirect.com/science/article/abs/pii/S0378778814006720

[7] S. Acha, Y. Du, N. Shah, Enhancing energy efficiency in supermarket refrigeration systems through a robust energy performance indicator, International Journal of Refrigeration 64 (2016) 40–50. doi:10.1016/J.IJREFRIG.2015.12.003.
URL https://www.sciencedirect.com/science/article/abs/pii/S0140700715003813

[8] S. Acha, A. Mariaud, N. Shah, C. N. Markides, Optimal design and operation of distributed low-carbon energy technologies in commercial buildings, Energy 142 (2018) 578–591. doi:10.1016/J.ENERGY.2017.10.066.
URL https://www.sciencedirect.com/science/article/pii/S0360544217317711

[9] V. Caritte, S. Acha, N. Shah, Enhancing Corporate Environmental Performance Through Reporting and Roadmaps, Business Strategy and the Environment 24 (5) (2015) 289–308. doi:10.1002/bse.1818.
URL https://onlinelibrary.wiley.com/doi/ 10.1002/bse.1818

[10] S. A. Kalogirou, M. Bojic, Artificial neural networks for the prediction of the energy consumption of a passive solar building, Energy 25 (5) (2000) 479–491. doi:10.1016/S0360-5442(99)00086-9.
URL https://www.sciencedirect.com/science/article/abs/pii/S0360544299000869

[11] I. Larivière, G. Lafrance, Modelling the electricity consumption of cities: effect of urban density, Energy Economics 21 (1) (1999) 53–66. doi:10.1016/S0140-9883(98)00007-3.
URL `https://www.sciencedirect.com/science/article/pii/S0140988398000073`

[12] G. Mavromatidis, S. Acha, N. Shah, Diagnostic tools of energy performance for supermarkets using Artificial Neural Network algorithms, Energy and Buildings 62 (2013) 304–314. doi:10.1016/J.ENBUILD.2013.03.020.
URL `https://www.sciencedirect.com/science/article/abs/pii/S0378778813001886`

[13] M. Yalcintas, Energy-savings predictions for building-equipment retrofits, Energy and Buildings 40 (12) (2008) 2111–2120. doi:10.1016/J.ENBUILD.2008.06.008.
URL `https://www.sciencedirect.com/science/article/abs/pii/S0378778808001357`

[14] J. Yang, H. Rivard, R. Zmeureanu, On-line building energy prediction using adaptive artificial neural networks, Energy and Buildings 37 (12) (2005) 1250–1259. doi:10.1016/J.ENBUILD.2005.02.005.
URL `https://www.sciencedirect.com/science/article/abs/pii/S0378778805000502`

[15] W. Chung, Y. Hui, Y. M. Lam, Benchmarking the energy efficiency of commercial buildings, Applied Energy 83 (1) (2006) 1–14. doi:10.1016/J.APENERGY.2004.11.003.
URL `https://www.sciencedirect.com/science/article/abs/pii/S0306261904002028`

[16] X. G. Casals, Analysis of building energy regulation and certification in Europe: Their role, limitations and differences, Energy and Buildings 38 (5) (2006) 381–392. doi:10.1016/J.ENBUILD.2005.05.004.
URL `https://www.sciencedirect.com/science/article/abs/pii/S0378778805000824`

[17] Z. Li, Y. Han, P. Xu, Methods for benchmarking building energy consumption against its past or intended perfor-

mance: An overview, Applied Energy 124 (2014) 325–334. doi:10.1016/J.APENERGY.2014.03.020.
URL https://www.sciencedirect.com/science/article/abs/pii/S0306261914002505

[18] R. Brown, C. Webber, J. Koomey, Status and future directions of the Energy Star program, Energy 27 (5) (2002) 505–520. doi:10.1016/S0360-5442(02)00004-X.
URL https://www.sciencedirect.com/science/article/abs/pii/S036054420200004X

[19] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047. doi:10.1016/J.RSER.2017.09.108.
URL https://www.sciencedirect.com/science/article/abs/pii/S136403211731362X

[20] J. Jeong, T. Hong, C. Ji, J. Kim, M. Lee, K. Jeong, Development of an integrated energy benchmark for a multi-family housing complex using district heating, Applied Energy 179 (2016) 1048–1061. doi:10.1016/j.apenergy.2016.07.086.
URL https://linkinghub.elsevier.com/retrieve/pii/S0306261916310273

[21] R. Arambula Lara, F. Cappelletti, P. Romagnoni, A. Gasparella, A. Lara, R. Arambula, International High Performance Buildings Conference. Paper 137. 3 rd International High Performance Buildings Conference at Purdue, Tech. rep. (2014).
URL http://docs.lib.purdue.edu/ihpbc/137

[22] M. M. Santamouris, Energy performance of residential buildings : a practical guide for energy rating and efficiency, James James/Earthscan, 2005.

[23] M. S. Spyrou, K. Shanks, M. J. Cook, J. Pitcher, R. Lee, An empirical study of electricity and gas demand drivers in large food retail buildings of a national organisation, Energy and Buildings 68 (2014) 172–182. doi:10.1016/j.enbuild.2013.09.015.

URL https://linkinghub.elsevier.com/retrieve/pii/S0378778813005914

[24] Y. Ding, X. Liu, A comparative analysis of data-driven methods in building energy benchmarking, Energy and Buildings 209 (2020) 109711. doi:10.1016/J.ENBUILD.2019.109711.
URL https://www.sciencedirect.com/science/article/abs/pii/S037877881932047X

[25] G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761–1768. doi:10.1016/J.ENERGY.2006.11.010.
URL https://www.sciencedirect.com/science/article/abs/pii/S0360544206003288

[26] M. Yalcintas, U. Aytun Ozturk, An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database, International Journal of Energy Research 31 (4) (2007) 412–421. doi:10.1002/er.1232.
URL https://onlinelibrary.wiley.com/doi/ 10.1002/er.1232

[27] Met Office 2020, Uk daily temperature data, uk daily humidity data , part of the met office integrated data archive system (midas). ncas british atmospheric data centre, [Accessed 09.10.2020].
URL http://catalogue.ceda.ac.uk/uuid/1bb479d3b1e38c339adb9c82c15579d8

[28] E. . I. S. Department for Business, Greenhouse gas reporting: conversion factors 2020 - GOV.UK (2020).
URL https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2020

[29] K. Kumar, G. S. M. Thakur, Advanced Applications of Neural Networks and Artificial Intelligence: A Review, International Journal of Information Technology and Computer Science 4 (6) (2012) 57–68. doi:10.5815/ijitcs.2012.06.08.
URL http://www.mecs-press.org/ijitcs/ijitcs-v4-n6/v4n6-8.html

[30] H. S. Park, M. Lee, H. Kang, T. Hong, J. Jeong, Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques, Applied Energy 173 (2016) 225–237. doi:10.1016/J.APENERGY.2016.04.035.
URL https://www.sciencedirect.com/science/article/abs/pii/S0306261916304834

[31] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347. doi:10.1093/bioinformatics/btq134.
URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq134

[32] R. Yokoyama, T. Wakui, R. Satake, Prediction of energy demands using neural network with model identification by global optimization, Energy Conversion and Management 50 (2) (2009) 319–327. doi:10.1016/J.ENCONMAN.2008.09.017.
URL https://www.sciencedirect.com/science/article/abs/pii/S0196890408003634

6. Acknowledgements

This research was supported by funds provided via the Imperial College London – Sainsbury's Supermarkets Ltd. Partnership.

Appendix A. Appendix

Appendix A.1. Variable Selection Results

Table A.4: Variables Removed from Models and Rationale for decision

Initial Variables	**Tech. factor**	**Low Var.**	**High Covar**	**RFECV**
Most Recent Investment Type	●	-	-	-
building has CHP	-	●	-	-
HDD	-	-	-	$*$
Humidity	-	-	-	$*$
Building Location Type	-	-	-	○
Type of Light Fixture	●	-	-	-
Last Update of Light fixtures	●	-	-	-
Building has Click and Collect	-	-	-	●
Building has Goods Online	-	●	-	-
Building has Clothing Section	-	-	●	-
Building has Petrol Station	-	●	-	-
Building has EV Charging	●	-	-	-
Building has Café	-	-	●	-
Number of additional shops in building	●	-	-	-
Refrigerant type	●	-	-	-
Rate of refrigerant lost	●	-	-	-
Solar panels Installed?	●	-	-	-

● - Removed ○ - Partially removed $*$ - Removed for electricity regression

Table A.5: Final Variables Used in Each Model

	MLR		**ANN**		**DT**	
Variable	**Elec**	**Gas**	**Elec**	**Gas**	**Elec**	**Gas**
Area	●	●	●	●	●	●
Op. Hours	●	●	●	●	●	●
Food Ratio	●	●	●	●	●	●
Building Age	●	●	●	●	●	●
Biomass?	●	●	●	●	●	●
Gas	●		●		●	
CDD	●	●	●	●	●	●
HDD		●	●	●	●	●
Humidity		●	●	●	●	●
Click and Collect		●	●	●	●	●
Town Edge	●	●	●	●	●	●
Town Centre	●	●	●	●	●	●
Retail Park		●	●	●	●	●
Small Town		●	●	●	●	●
Standalone		●	●	●	●	●
Suburban HS	●	●	●	●	●	●

● - Included

Appendix A.2. Sample Building Descriptions and additional performance discussion

Building F was identified as an extreme poor performer with respect to carbon residual. This was due to its use of a combined heat and power (CHP) technology, where power is generated through the building's heating mechanism. Although CHP lowered electricity demand, as suspected, more natural gas was used which was also considered more carbon intensive by each model. Therefore, all models classified this building in their worst 26% of performers. Focusing only on carbon emissions, CHP technology was no longer as sustainable as when such technology was implemented due to the decarbonisation of the UK's electricity supply. However, building F was still clustered into mid-low priority rankings, as gas was significantly cheaper than electricity. This illustrated the balance of financial and sustainability considerations weighed by the output clustering approach, such findings can enable nuanced investment decision-making.

Table A.6: Sample Building Details

Variable	A	B	C	D	E	F
Opening Year	2013	1980	1992	1978	1991	2000
Petrol Station	Yes	Yes	Yes	Yes	Yes	Yes
Charging Spaces	4	0	0	0	0	4
CHP Capacity	No	No	No	No	No	Yes
Weekly Opening Hours	101	101	101	101	142	96
Classification	Standalone	Edge of Town	Retail Park	Retail Park	Retail Park	Standalone
Cafe	Yes	Yes	Yes	Yes	Yes	Yes
Good Online	Yes	Yes	Yes	Yes	Yes	Yes
Click and Collect	No	No	Yes	No	Yes	Yes
PV Capacity	Yes	No	Yes	Yes	Yes	Yes
Gas Capacity	Yes	Yes	No	Yes	Yes	Yes
2019 Electricity CF	213.2	213.2	213.2	199.7	213.2	202.5
Food Ratio	0.164	0.411	0.376	0.353	0.426	0.049
Selection Reason	Leaking	High Demand	Leaking	Age	Largest Area	CHP

Appendix A.3. Variable Importance Variance Background

The below table categorised the variance in importance assigned to variables between periods by each model for both gas and electricity prediction.

Table A.7: Average Variance in Variable Importance Score by Period

Model	Electricity	Gas
	%	%
MLR	6.87	3.02
ANN	7.02	4.28
DT	0.05	2.58