Designing Sensor Sets for Capturing Energy Events in Buildings

Tom Lovett[a], JeeHang Lee[a,*], Elizabeth Gabe-Thomas[b], Sukumar Natarajan[a], Matthew Brown[c], Julian Padget[c], David Coley[a]

[a]*Department of Architecture and Civil Engineering*
University of Bath, Bath, UK, BA2 7AY
[b]*Department of Psychology*
University of Bath, Bath, UK, BA2 7AY
[c]*Department of Computer Science*
University of Bath, Bath, UK, BA2 7AY

Abstract

There is a growing desire to measure the operational performance of buildings – often many buildings simultaneously – but the cost of sensors and complexity of deployment is a significant constraint. In this paper, we present an approach to minimising the cost of sensing by recognising that researchers are often not interested in the raw data itself but rather some inferred performance metric (e.g. high CO_2 levels may indicate poor ventilation). We cast the problem as one of constrained optimisation – specifically, as a bounded knapsack problem (BKP) – to choose the best sensors for the set given each sensor's predictive value and cost. Training data is obtained from a field study comprising a wide range of possible sensors from which a minimum set can be extracted. We validate the method using reliable self-reported event diaries as a measure of actual performance. Results show that the method produces sensors sets that are good predictors of performance and the optimal sets vary substantially with the constraint parameters. Furthermore, valuable yet expensive sensors are often not chosen in the optimal set due to strong co-incidence of sensor signals. For example, light level and sound level often increase at the same time. The overall implication of the work is

*Corresponding author

Email addresses: t.r.lovett@bath.ac.uk (Tom Lovett), j.lee@bath.ac.uk (JeeHang Lee), e.g.thomas@bath.ac.uk (Elizabeth Gabe-Thomas), s.natarajan@bath.ac.uk (Sukumar Natarajan), m.brown@bath.ac.uk (Matthew Brown), j.a.padget@bath.ac.uk (Julian Padget), d.a.coley@bath.ac.uk (David Coley)

that a large number of co-incident low-cost sensors can be used to build up a picture of building performance, without significantly compromising information content, and this could have major benefits for the smart metering industry.

Keywords: Energy use, sensing, intelligence, interaction, ENLITEN

1. Introduction

The reduction of energy use in buildings has become a major challenge for researchers across multiple fields. The UK government has committed to 80% reductions in carbon emissions by 2020 [1], and a large proportion of these emissions stem from the operation and use of buildings [2]. Building energy efficiency aside, it is the occupants and their energy-related behaviour within the buildings that are a critical and complex factor in overall energy use [3].

To tackle the problem of energy usage reduction in buildings, researchers have used sensing technology to capture and analyse buildings' energy use so that efficiency can be improved and methods of lowering energy demand can be explored, e.g. through changing occupants' energy-related behaviour. The first step in enabling behavioural change is the gathering and sensing of pertinent data. As such, key questions emerge about how best to approach energy sensing: what sensors should we use? How many do we need? How intrusive and costly is the installation? Direct energy sensing with electricity and gas sensors is commonplace [4, 5, 6], but direct sensing alone does not account for total energy use, nor does it allow for non-trivial analyses of the often individualistic causal factors involved in energy consumption.

It is therefore important to look at the more abstract notion of energy events within buildings. Rather than monitoring how often a kettle is used, it may be more useful to monitor the events that involve kettle use, e.g. making breakfast, which could also comprise of other energy-consuming activities, e.g. using the hob or opening a window.

In order to be able to infer these events accurately, we need to capture the right data, which means that we need to deploy the right sensors in the right locations around the building. This alone is a non-trivial problem due to various factors such as health and safety, aesthetics (the best functional position for a sensor may not be the ideal position

aesthetically), power supply reach and – if the sensors are part of a sensor network – connectivity and range.

On top of this, there are cost installation issues. System designers often have fixed budgets and, if meaningful data is to be gathered, a sizeable number of buildings may need to be considered for sensor installation.

There are two key questions here: first, what are the "right" sensors for capturing energy events in a building, and how do we measure their value? Second, given this measure, what is the best sensor set for capturing such events in a building given certain constraints, e.g. budgetary and deployment constraints?

There are two key contributions in this paper:

- A method for assigning a value to a sensor in terms of its utility in capturing human activities that involve energy consumption in a building.
- A method for the selection of maximal value sensor sets subject to practical constraints such as budget and sensor quantities.

We compute a value metric for a given sensor in the context of a given deployment based upon a data set collected from a field study of domestic buildings in the UK. The study starts from the premise that by "over-sensing" a building, it becomes possible to identify the subset of readings, and thereby the sensors, that are necessary to capture the energy and occupant events that characterise the building's use. We encode a sensor value from an aggregate measure of feature value, as output by random forest feature selection methods. We then combine these values with monetary cost and model the resulting integer linear programming problem as a knapsack problem which, although NP hard, can be solved in pseudo-linear time ($\mathcal{O}(nW)$). We present some example sensor sets from our field study as budgetary and limit parameters vary, and illustrate how predictive certain sensors are – notably CO_2 and light level sensors – from others.

The outputs from this analysis allow the designers of energy sensing systems to determine the predictive values for each sensor in a candidate design set, and to choose sensor sets of maximal predictive value given budgetary and deployment constraints.

The rest of the paper is organised as follows: first, we review and contrast prior work in building energy sensing and sensor selection. Then we outline our high level

approach to sensor selection using random forests to estimate sensor value, and a bounded knapsack algorithm to choose the sensor sets based upon a range of constraints. We then describe our field study in UK homes before finally discussing the implications and limitations of the presented work.

2. Related Work

2.1. Capturing Energy Events in Buildings

The use of technology to sense, infer and predict energy use in buildings has become increasingly popular as demand for energy efficiency rises. As such, it is a broad field, with different disciplines focusing on many areas of energy use in buildings, from appliance and HVAC usage [7, 8] to occupants' behaviour [9] and responses to energy feedback [10, 11].

There is a recognised strong correspondence between the actions of building occupants and energy use [9]. As a consequence, there has been a focus on occupant activity recognition in relation to monitoring energy consumption and improving energy efficiency. Much indoor activity recognition is concerned with the inference of general activities, e.g. whether the occupants are sleeping, but our objective is capturing particular activities that consume energy. Prior work in this area ranges from direct sensing to higher level inference and automation [12]. In [13], Milenkovic and Amft focus on energy activities in an office space. By using a hidden Markov model (HMM) which received inputs from passive infra-red (PIR) motion sensor, they were able to predict desk-based work to a high degree of accuracy, with simulation results predicting $\approx$ 20% energy savings if control systems used these data. Similarly, PIR sensors are used for improving energy management through occupancy classification by Agarwal *et al.* [12] who, through simulation, show that potential energy savings of up to 15% may be achieved by integrating occupancy detection into building energy management systems.

Depsite the correlation between energy use and occupant action [9], much of the literature focuses on occupancy detection with hardly any consideration of the occupants' effect on energy use. Studies into occupancy detection do tend to cite energy efficiency

as a motivating application, but concentrate on the performance of the occupancy detector [14]. In [15], Patel *et al.* use HVAC air pressure sensors to infer occupancy as well as door and window opening/closing events. Notable domestic sensing work that focuses more on energy use rather than occupancy includes Cohn *et al.*'s GasSense [4] – which uses the sound of domestic gas relief valves to measure gas events in the home – Gupta *et al.*'s ElectriSense [16] – which uses electromagnetic interference (EMI) signatures to monitor appliance electricity use – and Froehlich *et al.*'s HydroSense [17], which classifies water usage events through pressure changes.

Our work focuses on more than just occupancy detection; rather we concentrate on sensing *energy events* i.e. human activity involving energy consumption. Similar studies tend to focus on atomic energy events, e.g. what appliances are being used [5], but we consider more abstract events such as "preparing food", which can incorporate multiple atomic events; often concurrently. This aligns with the idea that occupant behaviours have strong relationships with energy efficiency [3].

Attempting to recognise more abstract events comprised of multiple directly detectable events is an approach that has been used previously by Wilke *et al.* [18] to model real-time occupancy in buildings. Our work is similar, in that we too use the Multinational Time Use Survey to determine interesting events [19]. Again, however, our focus is on events that consume energy, rather than those that determine occupancy.

There is an increasing industrial demand for energy sensing and occupant behaviour learning with a view to saving energy. Commercial systems such as NEST[1] – which uses a variety of environmental sensors – are popular, although they do require occupant training and the intelligent features have suffered from usability issues [6].

Finally, energy sensing in buildings is typically performed through direct sensing of electricity use through whole-building and plug-load electrical sensing [7], disaggregation of appliance use from electrical sensing traces [5, 16] and direct gas use sensing [4]. However, comparatively few studies have considered deploying environmental sensors to infer energy use; mainly because these sensors are not designed for direct energy measurement. There is potential predictive value in using environmental

[1] `http://www.nest.com`

sensors in conjunction with direct energy sensors, and our work in this paper concentrates on measuring this predictive value prior to selecting the appropriate sensors for the application.

2.2. Sensor Selection Approaches

The goal of sensor selection is to choose from an existing set of sensor inputs in order to maximise some objective function or parameter [20, 21]. Part of our contribution in this paper is the derivation of sensor "value" in terms of its utility in capturing energy events. In contrast to other sensor selection approaches, we are concerned with the more practical problem of sensor selection *a priori*, i.e. choosing the sensor set design prior to deployment in an application given the practical constraints in doing so, rather than choosing the best measurement from a pool of existing sensors.

The closest work to the study in this paper is Zhang *et al.*'s study of feature selection for occupancy classification in office spaces [22]. Here, the authors explore the relative information gain – or uncertainty coefficients – as a value measure for a small range of sensors using intermittent ground truth gathered in an office environment. We use a different measure of sensor value in a domestic environment, but our results broadly support Zhang *et al.*'s, which show that sound and CO_2 sensors appear to be the most effective at detection; albeit for energy events in ours, and occupancy events in theirs. By incorporating sensor costs, however, we show that these sensors are not always the best ones to choose for maximising sensor value given a set of constraints. In brief, the proposal in [22] pursues "the best sensor sets" for the occupancy detection in an office setting, based only on sensor values, and without consideration of the constraints (e.g. each sensor cost which affects the total cost of the sensor sets). Our approach instead attempts to consider the combination of sensor values, constraints, and sensor redundancy [2] in order to find out the best sensor set that meets the cost constraints. For example, a CO2 sensor could be the most significant for the desired detection (for both energy event and occupancy), but the unit cost is relatively high. If we have a cost constraint on the final choice, our approach could suggest other alternatives, by consid-

Response to R3

[2] See Section 3.5 for more details.

ering both the cost and the efficiency, such that when the cost of a CO2 sensor exceeds the available budget, a cheaper sensor or sensor set (e.g. temperature and PIR sensors) could be suggested as an alternative proposal that satisfies the cost constraint as well as efficiency of energy event detection.

Our use of knapsack algorithms, or integer linear programming in general, is not new, although the application to sensor selection for energy event capture, as far as we are aware, is. The use of knapsack algorithms has previously been applied to the domain of sensing, typically for time-dependent resource usage. In [21], Joshi and Boyd use convex optimisation to develop a heuristic approach that approximates sensor subsets for minimising the error of parameter estimation. Godrich *et al.* directly use the knapsack problem to formulate optimal configurations for radar architectures [23], and Bian *et al.*[20] use a more general form of linear programming to select a subset of sensors from a theoretical global set based upon maximum utility. Here, utility is somewhat abstract, although the authors do give an example of expected variance reduction in average sensor measurements, i.e. the usefulness of a sensor is its accuracy.

In summary, our work seeks to aid in the *a priori* choice of sensors for capturing energy events in buildings. By combining work in sensor selection problems with the field of energy sensing, we present an original design approach.

3. Sensor Selection and Study Design

In this section, we describe our method for designing sensor sets to capture energy events in buildings. We first define the problem statement in greater detail, before describing the general design approach of assigning a value measure to each sensor and choosing sets using a BKP algorithm. We also outline our means of measuring sensor redundancy, i.e. the amount each sensor can be predicted from others, and detail the methodology of our field study.

3.1. Problem Statement

Our general problem statement is: what is the best sensor set design for capturing energy events in a particular building? The "best" sensor set needs a more concrete

definition however, and the best set is unlikely to be the same across all building types. The best set for a single building with stringent accuracy requirements will not be the best set for a large deployment with limited budgetary requirements. Rather than define a global "best set", we give an approach to determining the best set given contextual parameters, e.g. budget and scale of deployment.

Thus, we refine the problem statement to be: what then is the best sensor set design for capturing energy events in a particular building *for a given set of parameters*? In this case, the best set is one that maximises the information required for event capture, whilst meeting the cost requirements of deployment. We can set this up as a constrained optimisation problem, which requires that each sensor have a measure of cost and value, and solves the maximum value achievable given the cost constraints. With these cost and value measures, the constrained optimisation problem becomes a form of the famous *knapsack problem* [24], which can be solved in pseudo-linear time using dynamic programming.

Thus, the key problem is not so much the optimisation process, but the *determination of sensor cost and value*. Cost may typically be simply defined as the financial cost (but see §3.4 for discussion of other factors), so it is *sensor value* that is the key measure to define. In the next section, we formally outline the constrained optimisation problem, before detailing our approach for calculating sensor values.

3.2. Constrained Optimisation: The Knapsack Problem

The knapsack problem is a simple integer linear program that seeks to find the optimal combination of n distinct items that maximises the total value of a weight-constrained knapsack, given that each item has a value and a weight. More formally, given n distinct items, where each item i has a corresponding value v_i, number of copies x_i and weight w_i, and an overall weight constraint W, the knapsack problem seeks to:

$$\text{maximise:} \quad \sum_{i=1}^{n} v_i x_i \tag{1}$$

$$\text{subject to:} \quad \sum_{i=1}^{n} w_i x_i \leq W \qquad (2)$$
$$x_i \in \{0, \ldots, c_i\}$$

where c_i is an upper bound on the number of copies of each item. c_i could be viewed as a sensor quantity limit, e.g. a stock limit. The above problem is a bounded knapsack problem (BKP), which does not restrict the items in the knapsack to one copy each; as is the case for the 0-1 knapsack problem (KP). The BKP can be solved by reduction to a KP, allowing a dynamic programming solution in $\mathcal{O}(nW \log W)$ [24] or $\mathcal{O}(nW)$ [25].

Thus, we can apply the knapsack problem to the problem of designing sensor sets for energy event capture in buildings. Instead of items, we have sensors with a measure of predictive value for capturing energy events, and instead of weights, we have a measure of cost.Within this context: (i) n distinct items correspond to a number of distinct sensors that our study considers (e.g. Temperature, Humidity, Light, Sound, CO2 etc), (ii) each item i corresponds to each sensor with v_i denoting the predictive value of the sensor i, (iii) number of copies x_i is a quantity of the sensor i, (iv) the weight w_i corresponds to the (financial) cost of the sensor i, and (v) the overall weight constraint W corresponds to the budget i.e cost requirements of deployment. Our final knapsack is the chosen sensor set that maximises Equation 1 with the number of sensors (x_i) and their predictive value (v_i), and the weight constraint (W) is a budget over some cost measure which is specified in Equation 2.

Response to R1

Thus, for each sensor, we need to determine:

- **Value (v_i):** A measure of each sensor's value, in terms of its response to energy events in the building.
- **Cost (w_i):** An applicable measure of cost, or "weight" in the knapsack problem.

Given values for v_i and w_i, a quatity x_i of the sensor i can be installed at x_i locations in the domestic buildings.

3.3. Defining Sensor Values

Determining a measure of value for a sensor is context-dependent and potentially non-trivial. In our case, a more valuable sensor provides better information about energy events in a building than a less valuable one. For each sensor, we define a number of features of its raw measurement, and view the problem as a *feature selection* problem, i.e. what sensor features are better predictors of energy events in buildings. We can then aggregate each feature's value measure into an overall value measure for the sensor.

3.3.1. Feature Extraction

Before undertaking feature selection, we must define and calculate the sensor features that we wish to measure through *feature extraction*. This is done because we believe that some feature such as first-order difference in sensor measurements or moving average of sensor measurements will be more strongly predictive of energy events than the raw measurement alone. The definition of a feature is a free choice for the designer, and there is no limit to the type or number of features that can be chosen for feature extraction. Again, this is likely to be context dependent, and we define the features for our field study in §3.7.

3.3.2. Feature Selection: Random Forest

To perform feature selection, we use a random forest process on the extracted features. A random forest is an ensemble method that combines a set of decision tree classifiers, each of which is comprised of a random sample of input variables (in our case, extracted features). For brevity, we refer the reader to Breiman's description of the random forest method for a detailed overview [26]. We use random forests to measure the value of each extracted sensor feature using the average decrease in node impurities from splitting the decision trees on that feature.

For this, we use the Gini impurity measure, i.e. the greater the *decrease* in the Gini impurity for the feature variable – averaged over the forest – the more important the feature variable. Thus, to measure the value of each sensor, we uses the mean Gini impurity decrease over the features attributable to each sensor, since the inclusion or

exclusion of a sensor adds or removes its entire feature set. Moreover, sensor values are unlikely to be independent, and the mean Gini decrease provides a way to average the incremental effect of each sensor in the candidate set. Thus, we use mean Gini decrease over the sensor's feature set as the sensor's value measure in the knapsack problem.

3.4. Defining Sensor Costs

As with the choice of value measure, the choice of cost measure is likely to be context-dependent. An obvious choice is the financial cost of each sensor, but more complex cost functions could be designed that incorporate, for example, sensor energy costs, installation effort or sensor reliabilities. In addition to budgetary constraints, logical constraints can be introduced that restrict the chosen sensor set to particular subsets of the overall power set (all 2^n possible choices of sensor set from n sensors).

3.5. Sensor Redundancy

Once a sensor set is found according to defined sensor costs and values, design decisions surrounding the pruning of sensors may be aided through measuring *sensor redundancy*, i.e. how much information about a sensor's value can be predicted from the others in the set? In [22], Zhang *et al.* use an information theoretic approach to select features for occupancy detection using environmental sensors in an office, and we use a similar approach here for energy events [3]. Using the entropy function from information theory for each sensor output:

Response to R1 and R3

$$H(X) = \sum_{x \in X} p(x) \log \left(\frac{1}{p(x)} \right) \tag{3}$$

where $H(X)$ is an entropy, x is a random variable, and $p(x)$ is the probability of $X = x$. In this work, the random variable x corresponds to a sensor feature measured and derived from one sensor.

Response to R1

[3] Zhang *et al.* use information theory to study the correlation between occupancy levels and features extracted from various environmental measurements [22]. Conversely, we use information theory for the purpose of identifying the correlation between each pair of sensors in the set i.e. how much information about the feature extracted from measurements of one sensor can be predicted by that of the other.

$I(X;Y)$ is the mutual information content of variables X and Y:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{4}$$

where x and y are random variables, $p(x)$ and $p(y)$ are the probability of $X = x$ and $Y = y$, respectively, and $p(x,y)$ is the conditional probability of x given y. As noted previously, the random variable $x \in X$ corresponds to the feature of one sensor and the random variable $y \in Y$ corresponds to that of the other. Thus, $I(X;Y)$, is the *mutual information content (MIC)* between two co-present sensors, X and Y, and is a measure of the quantity of common information that can be derived from them.

Response to R1

Then, calculate the uncertainty coefficient:

$$C_{XY} = \frac{I(X;Y)}{H(Y)} \tag{5}$$

That is, the proportion of bits about sensor feature Y that can be predicted from sensor feature X.

We calculate the uncertainty coefficient C_{XY} over all sensor feature pairs $X \times Y$ to explore possible redundancy in sensor set selections. That is, once a sensor set is chosen, one can use the uncertainty coefficient measures to remove further sensors from the set if needs be. For example, let us assume that the uncertainty coefficient between the temperature and the CO_2 sensor is high. This suggests that there is high probability that the information measured and derived from the CO_2 sensor can be predicted by the temperature sensor, which is to say that the CO_2 sensor is redundant in the presence of a temperature sensor. If the cost of CO_2 sensor is much greater than a temperature sensor, then the sensor set designer may be able to choose a temperature sensor instead of a CO_2 sensor for energy event capture. This is not to suggest that a temperature sensor would *replace* the CO_2 sensor for physical measurements, but rather that when the intention is to detect changes in signals as a proxy for occupancy then either sensor is likely to provide the same signal at the same strength. This could be done before the sensor set is chosen, but it would be prudent to observe the sensor's influence in the chosen set before considering pruning it based upon redundancy. In our field study, we

Response to R1

present the uncertainty coefficients for all co-present sensors (sensors in the same room of each home) to illustrate the redundancy in our buildings' sensors.

3.6. Field Study

In order to demonstrate how a sensor set for capturing energy events can be chosen, we present the results of a field study in a set of domestic buildings in the UK. We recruited 4 homes to be studied for the duration of 7 consecutive days in August 2013. Details of the homes including house type, number of bedrooms and number of occupants are summarised in Table 1. The experimental settings including the sensor placement in each home are shown in Figure 1[4]. Within certain rooms in each home – each room common to each home – we installed the following sensors:

Response to R2

- **Kitchen:** Temperature, light, humidity, PIR, CO_2 and sound level sensors.
- **Living Room and study:** Temperature, light, humidity, PIR, and sound level sensors.
- **Main bedroom and secondary bedroom:** Temperature, PIR and sound level sensors.

Temperature was recorded in °C, light in lux, CO_2 in ppm, motion in $\{0, 1\}$ and sound level in dB. Each of the room's sensors were connected to a single Arduino Uno

[4]Note that the layouts are not exactly the same as the participants' dwellings, but equivalent to the building details they provided. We use these examples to show the context for the sensors for the field study.

Home	Type	Bedrooms	Floors	Occupants
A	Terraced	3	3	2
B	Semi-detached	3	2	3
C	Detached	4	2	3
D	Terraced	4	3	2

Table 1: Descriptions of the homes used in the field study.

board, (5 boards per home) which was housed in an acrylic plastic box shown in Figure 2. The sensors were placed on surfaces such as bookshelves and kitchen counters, and each of them sampled data at a rate of once per minute. The data were sent to us remotely over the home's WiFi connection and simultaneously logged locally to an SD card in order to reduce risk of data loss.

To capture a record of ground truth events in each home, we asked the primary occupant to record energy-related events around the home throughout the week in a diary study. This was considered appropriate over ethnographic methods as it allows examination of temporal sequences across an extended time period in a practical and accessible manner [27].

To define the energy events, we used Oxford University's Multinational Time Use Study (MTUS) data [19], selecting domestic event codes that classify energy-consuming events around the home, similar to a method used by Wilke *et al.* [18] to predict building occupant activities.

The primary occupant was presented with the list of events in Table 2 as guidance on the type of events to capture. Then, throughout the duration of the study, the occupant was asked to log as many of them as possible in Google's calendar application so that we could capture the event description, its location, i.e. room, and the start and end times of the event. The occupants were given no restriction on the event description text, i.e. although the MTUS data was used as a guide to the *type* of events to capture, participants were free to use their own label descriptors.

In addition to these events, participants were asked to record known periods where homes were unoccupied and no energy events were undertaken. This allowed to us encode ground truth for each room into a variable with three levels:

- Known energy event: the occupant logged an energy event.
- Known absence of event: the occupant logged an absence of energy events.
- Nothing recorded: the occupant did not log anything, i.e. ground truth is unknown.

We dismissed data during which the occupants did not log anything, i.e. the ground

truth was unknown. Although this reduces the size of the dataset for analysis, it is a manifestation of using self-report methods to capture ground truth. Participants are unlikely to capture everything, but their behaviour is perhaps more "natural" than if other methods, e.g. ethnography, were used. The diary study also minimises the risk of retrospective bias common to other self-report methodology as the recorded events were objective and concrete by nature [28]. Furthermore, the neutrality of the events recorded should minimise social desirability bias.

Ethnographic methods are also time consuming for both the researcher and the participant, which may compromise on both study validity and the duration of the data capture. We attempted to minimise participant burden further through the presentation of clearly defined event classes (Table 2) [28].

3.7. Extracted Features

For each of the sensors, we calculated the following features:

- Raw value at timestep k: y_k
- First order difference: $\Delta(y_k) = y_{k+1} - y_k$
- Second order difference: $\Delta^2(y_k) = \Delta(y_{k+1}) - \Delta(y_k)$
- Simple moving average, over a m minute window:

$$\bar{y}_k = \frac{1}{m} \sum_{i=k-m+1}^{k} y_i \tag{6}$$

These are similar features to those used by Zhang *et al.* in [22] in their study of sensor feature selection for office space occupancy detection.

3.8. Section Summary

In summary, we have outlined our method of sensor selection using a BKP algorithm. We have also described our measure of a sensor's value in terms of its utility in capturing energy events in buildings, as generated from random forests. Furthermore, we have defined a measure of redundancy within a sensor set using the uncertainty coefficient. Following the methodology of our field study, the next section presents the results of applying the techniques in this section to the data obtained from the study.

4. Results

This section presents the results from our field study and uses the procedure outlined in § 3 to identify the best sensor set. We first show the sensor values calculated using the random forest approach, along with the observed sensor redundancies as measured by the uncertainty coefficient in Equation 5. We then examine various example sensor sets output from the BKP algorithm using these sensor values and a list of illustrative costs.

4.1. Configuration Parameters

All sensor data were captured at a sample rate of once per minute. For the random forest process, our study dataset is split .7 training data, .15 validation data and .15 test data. Each forest consists of 500 trees, with 4 variables randomly sampled per split; no replacement. We used the *R* package "randomForest" [29] to run the random forest process with the aforementioned parameters. This package uses Breiman's approach [26].

For the BKP, we use Pferschy's $\mathcal{O}(nW)$ BKP algorithm described in [25]. For the probability distributions $p(x)$ in Equation 3, we use implicit probability estimators from the dataset frequencies. For the moving average feature, we set m – the moving average window – to 20 minutes for each sensor. The sensor values for the BKP are set to the mean Gini decrease measures for each sensor. For the sensor costs, we use the approximate financial cost of the sensors in our study setup, which includes the cost of each sensor itself plus a portion of the hardware required to acquire data from it remotely, e.g. CPUs and WiFi hardware. We must stress that this measure is illustrative for the purposes of demonstrating our sensor selection process, and should *not* be viewed as a standalone measure (unlike the sensor value measure) – the costs are financially realistic at the time of writing, but obviously varies across manufacturers, suppliers, time and market. The costs for the sensors are as follows: 215 for CO_2, 20 for humidity, 16 for light, 115 for sound and 17 for temperature.

Figure 3 shows a plot of the raw sensor data from Home 1's kitchen over the duration of the study. Figure 4 is an example of an energy event recorded on Aug 06,

Response to R2

2013 by Home 1, retrieved from Google's calendar application. These participant-recorded energy events are encoded into three 'event' states seen at the bottom row, entitled 'Event', in Figure 3: (i) 1 corresponds to a participant-recorded energy event in this particular room, a kitchen (e.g. Food in Kitchen, Coffee in Kitchen, Cooking in Kitchen, Dishwasher in Kitchen), (ii) 0 corresponds to a participant-recorded "non-event" and (iii) NA corresponds to no record. Here is an example. As seen in Figure 4, 5 energy events are recorded by Home 1's participant. Since there is no PIR event observed in the morning, only 4 events in the afternoon are encoded as 1 representing an energy event in the kitchen on Aug 06, 2013.

The study participants logged 392 events in total over the 7 days (A = 119, B = 59, C = 77, D = 137).

4.2. Sensor Values

Figure 5 shows the top 10 ranked feature set as output from the random forest process using the mean Gini impurity decrease as a value measure. Figure 6 shows the mean Gini impurity decrease for each sensor, averaged over the sensor's features.

Figure 7 shows the uncertainty coefficients of each sensor's raw measurement relative to the others, i.e. the approximate proportion of bits that can be predicted about sensor j from sensor i. Note, this is only calculated using sensors that are co-present, i.e. sensors that are located on the same Arduino board in the same room of each study home.

4.3. Optimal Sensor Sets

Figure 8 shows a set of example sensor sets output by the BKP algorithm for given weight constraints (W) and upper bounds on the sensor quantities c_i. The values are the mean Gini impurity decrease measures in Figure 6, and the costs are described in Section 4.1 above.

5. Discussion

This section discusses the results and their implications and limitations for energy sensing in buildings. The two key outputs from our work are (i) a quantitative measure

of sensor "value" as a predictor of energy events; and (ii) an approach for designing sensor sets for energy sensing in buildings based upon values and a measure of cost.

5.1. Implications

The first implication of this work relates to the utilisation of environmental sensors as predictors of energy events in buildings. The sensors in our study are designed to measure a particular environmental property, e.g. temperature, rather than direct energy use – something that devices such as current clamps attached to electricity meters and plug power monitors do. The sensor values show that temperature, humidity, light, CO_2, sound and motion sensors are useful predictors of energy use, though their predictive values do vary both across sensors and between homes.

By combining these values with costs – in our case, financial costs – it is interesting to note that some of the more valuable sensors, e.g. CO_2 and sound, are not often included in the design sets output by the BKP solver (see Figure 8). Clearly this is because the building's sensing value can be maximised by using multiple low-cost, less valuable sensors rather than fewer high-cost, more valuable ones.

Other interesting results include the comparatively low Gini measure for the PIR motion sensor. Although, from Figure 3, motion appears to visually correspond to energy events, it is an event-based sensor and even its moving average value is not an outstanding predictive feature. There are also issues relating to stationary people not triggering the sensor, and the argument that a motion sensor is not a presence sensor [13]. A probabilistic input such as a pre-learned HMM may be more suitable to increase this value. Despite this however, the PIR sensor tends to be chosen for mid-budget sensor sets due to its low cost.

From Figure 7, we see that CO_2 shares an almost uniform amount of information with the other displayed sensors and that light level shares the largest (in mean value). The CO_2 result broadly agrees with the value from the office study in [22], although humidity is lower. A higher coefficient implies redundancy in the sensor information, which could be used by the designer to prune sensors from the set if necessary. The uniform – and relatively high – uncertainty coefficient for the CO_2 sensor, coupled with its typically large financial cost stands in contrast to its large though variable sensor

value (see Figure 6).

This work has further implications for designers of energy sensor systems. By choosing the sensors *a priori*, deployment costs can be saved by lowering sensor redundancy, though it is probably wise to test a larger set in a pilot study as we have done here. Although our sensor values can be taken as a measure of predictive value, this value is likely to be context specific, i.e. our field study was conducted in domestic buildings, and we recommend that designers replicate our approach in order to obtain customised sensor values. However, the values presented in the results can be used as a guideline to the predictive power of the sensors in a domestic context.

There is also an interesting argument for using a KP solver rather than a BKP one (as we have used in this paper) for the sensor set specification. The BKP solver allows multiple copies of each sensor to be included in the final building set; therefore the physical sensor units, e.g. the Arduino or Raspberry Pi extension boards, may vary in their design in order to accommodate multiple sensors in different locations. By using a KP solver, a single, consistent sensor unit can be designed that only allows one copy of each item in the output set. The advantage here lies in the parsimony of general design, but it does restrict the amount of energy information that could be extracted from a building compared with a BKP set. Thus, there is a design trade-off between simplicity and value that the designer should make. It is relatively trivial to run a KP solver using the process in this paper, so the output sets can be compared without much further work.

Scalability is another key implication of our work. As sensors vary in cost and budgets are typically fixed, designers and researchers may face the problem of choosing a large sensor set for a small number of buildings, or a sparser sensor set for a larger number of buildings. Using our approach, these constraints can be fixed – see the examples in Figure 8 – to suit the design requirements. Likewise, if there are sensors that are essential to the application requirements, they can be removed from the candidate set and the BKP may be run on the remaining set.

Finally, our approach can be generalised beyond domestic buildings. Although our field study was conducted within the home, there is no restriction to this, but we do suggest that new sensor values be derived for environments other than domestic ones.

Furthermore, various value and cost functions may be used. In this paper, we have used the Gini impurity measure for value, and approximate financial cost for the cost measure. Again, there are no restrictions to the measures used – particularly for cost – as functions could be designed to combine, for example, financial cost with energy or installation disruption costs.

5.2. Limitations

The main limitations of our work relate to the context of sensing, the range of sensors and the study size. As discussed in the previous section, the context of sensing is important and the results obtained here are more applicable to, though not restricted to, domestic buildings. Furthermore, our range of sensors could be extended, as well as the features chosen for analysis in the random forest process.

Indeed, there are many parameters to explore in the feature extraction step. In addition to the choice of features, parameters such as moving average type or time window can be varied.

Other means of defining sensor values could also be derived. We chose to use the random forests approach due to its robustness and frequent use in the feature selection problem [26], but other approaches incorporating dimensionality reduction, e.g. principal components analysis (PCA), or regression models, e.g. generalised linear models (GLMs) or partial least squares analysis, could be used instead.

As we previously mentioned, financial cost is used in this paper as an illustrative cost measure, but other costs could be defined that incorporate, for example, installation effort or sensor energy use. Furthermore, the BKP algorithm is very sensitive to the cost and value measures, thus a robust measure of each would be useful for future work.

Our range of sensors was relatively small, and large projects are likely to consider a greater range than the environmental ones used in our study. However, this does not detract from the generalisability of our approach: random forests and the BKP solver can handle larger inputs.

Finally, our study was also comparatively small due to the constraints of fine-grained ground truth collection. As discussed in § 3, ground truth collection is laborious, and alternatives to the diary study are likely to compromise on data validity

[28]. Using a larger dataset gathered from more homes would reduce the uncertainty in the value measures in Figure 6; in particular, even though CO_2 and sound sensors have the larger mean Gini decreases, they also have the largest variances observed in the study data, thus more data would reduce this variance to give more accurate empirical measures of sensor value.

5.3. ENLITEN Deployment

We have used the process outlined in this paper to design a sensor set on the ENLITEN project [30], which aims to sense a wide range of energy-related, environmental and occupancy properties for domestic energy reduction. Along with direct energy sensors such as current clamps, gas meters and plug-load monitors, we have used the design process in this paper to create cheap, wireless sensor units from Raspberry Pi computers. At the time of writing, Raspberry Pis are inexpensive computing devices with standard hardware interfaces such as USB and Ethernet. They run a small operating system, and also contain a general purpose hardware interface.

Figure 9 shows a Raspberry Pi computer with our custom board containing the sensors output from the BKP solver: temperature, humidity, light and motion. We are deploying three of these sensor units per home (with a target deployment of 200 homes), with another temperature-only unit for monitoring radiator and boiler temperatures. All sensor units report their data in real time over WiFi through the occupants' broadband connection or a mobile data connection.

6. Conclusion and Future Work

In this paper, we have presented a process for designing sensor sets to capture energy events in buildings. The key contributions lie in the use of random forests to produce a measure of sensor value *a priori*, and the implementation of a bounded knapsack problem (BKP) solver that chooses an optimum sensor set given a set of costs and values. Through a field study in 4 UK homes, we have illustrated how random forests can be used to output a measure of predictive value using the Gini impurity measure, and how this measure – when combined with an appropriate cost measure,

e.g. financial cost – can be used to generate sensor sets given designer constraints. Through this, we have also shown that more valuable but expensive sensors such as CO_2 are often not included in the sets due to their high cost. Furthermore, we have shown that CO_2 and light sensors are particularly predictable, with a mean predictable proportion for both of ≥ 0.4 bits from the other sensors used in our study of domestic buildings (temperature, humidity and sound level).

For future work, we suggest replicating our field study in other building types, e.g. industrial buildings, and comparing further measures of cost beyond the purely financial. As we are currently deploying our sensor sets in the ENLITEN project, a large part of our future work involves validating how well the sensors perform as inputs to building energy and occupancy model. Another potential research is a comparison of real data collected by our sensor sets and simulation results e.g. to analyse the uncertainly coefficient matrix for co-present sensors with pre-simulation for the environment parameters, which allows the evaluation of the approach we present e.g. the information theoretic approach in measuring sensor redundancy.

Acknowledgments

This work is funded by EPSRC grant reference EP/K002724/1. All data created during this research are openly available from the University of Bath data archive at .

References

[1] Climate Change Act, retrieved from `http://www.legislation.gov.uk/ukpga/2008`. Dec 2013 (2008).

[2] Committee on Climate Change. Meeting Carbon Budgets – ensuring a low-carbon recovery, retrieved from `http://www.theccc.org.uk/reports`. Dec 2013 (June 2010).

[3] H. Allcott, S. Mullainathan, Behavioral science and energy policy, Science 327 (5970) (2010) 1204–1205.

[4] G. Cohn, S. Gupta, J. Froehlich, E. Larson, S. N. Patel, GasSense: Appliance-level, single-point sensing of gas activity in the home, in: Proc. Pervasive '10, Springer, 2010, pp. 265–282.

[5] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. Reynolds, S. Patel, Disaggregated end-use energy sensing for the smart grid, Pervasive Computing, IEEE 10 (1) (2011) 28–39.

[6] R. Yang, M. W. Newman, Learning from a learning thermostat: lessons for intelligent systems for the home, in: Proc. Ubicomp '13, ACM, 2013, pp. 93–102.

[7] J. Kolter, M. Johnson, REDD: A public data set for energy disaggregation research, in: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, 2011.

[8] O. Parson, S. Ghosh, M. Weal, A. Rogers, Nonintrusive load monitoring using prior models of general appliance types, in: Proc. AAAI '12, 2012.

[9] Z. Yu, B. Fung, F. Haghighat, H. Yoshino, E. Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, Energy and Buildings 43 (6) (2011) 1409–1417.

[10] E. Costanza, S. Ramchurn, N. Jennings, Understanding Domestic Energy Consumption through Interactive Visualisation: a Field Study, in: Proc. Ubicomp '12, 2012, pp. 216–225.

[11] S. Darby, The Effectiveness of Feedback on Energy Consumption, Working Paper, Oxford Environmental Change Institue (2006).

[12] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, T. Weng, Occupancy-driven energy management for smart building automation, in: Proc. 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, ACM, 2010, pp. 1–6.

[13] M. Milenkovic, O. Amft, An opportunistic activity-sensing approach to save energy in office buildings, in: Proc. e-Energy '13, ACM, 2013, pp. 247–258.

[14] V. L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A. E. Cerpa, M. D. Sohn, S. Narayanan, Energy efficient building environment control strategies using real-time occupancy measurements, in: Proc. 1st ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, ACM, 2009, pp. 19–24.

[15] S. N. Patel, M. S. Reynolds, G. D. Abowd, Detecting human movement by differential air pressure sensing in HVAC system ductwork: An exploration in infrastructure mediated sensing, in: Proc. Pervasive '08, Springer, 2008, pp. 1–18.

[16] S. Gupta, M. S. Reynolds, S. N. Patel, ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home, in: Proc. Ubicomp '10, ACM, 2010, pp. 139–148.

[17] J. E. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, S. N. Patel, HydroSense: infrastructure-mediated single-point sensing of whole-home water activity, in: Proc. Ubicomp '09, ACM, 2009, pp. 235–244.

[18] U. Wilke, F. Haldi, J.-L. Scartezzini, D. Robinson, A bottom-up stochastic model to predict building occupants' time-dependent activities, Building and Environment 60 (2013) 254–264.

[19] K. Fisher, J. Gershuny, Multinational Time Use Study. Chapter 3: Activity Codes, retrieved from `http://www.timeuse.org/sites/ctur/files/858/mtus-user-guide-chapter-3.pdf`. Dec 2013 (July 2013).

[20] F. Bian, D. Kempe, R. Govindan, Utility based sensor selection, in: Proc. IPSN '06, ACM, 2006, pp. 11–18.

[21] S. Joshi, S. Boyd, Sensor selection via convex optimization, Signal Processing, IEEE Transactions on 57 (2) (2009) 451–462.

[22] R. Zhang, K. P. Lam, Y.-S. Chiou, B. Dong, Information-theoretic environment features selection for occupancy detection in open office spaces, Building Simulation 5 (2) (2012) 179–188.

[23] H. Godrich, A. P. Petropulu, H. V. Poor, Sensor selection in distributed multiple-radar architectures for localization: A knapsack problem formulation, Signal Processing, IEEE Transactions on 60 (1) (2012) 247–260.

[24] S. Martello, P. Toth, Knapsack problems: algorithms and computer implementations, John Wiley & Sons, Inc., 1990.

[25] U. Pferschy, Dynamic programming revisited: improving knapsack algorithms, Computing 63 (4) (1999) 419–430.

[26] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32.

[27] K. Ellegård, J. Palm, Visualizing energy consumption activities as a tool for making everyday life more sustainable, Applied Energy 88 (5) (2011) 1920–1926.

[28] N. Bolger, A. Davis, E. Rafaeli, Diary methods: Capturing life as it is lived, Annual Review of Psychology 54 (1) (2003) 579–616.

[29] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (3) (2002) 18–22.

[30] T. Lovett, E. Gabe-Thomas, S. Natarajan, E. O'Neill, J. Padget, 'Just enough' sensing to ENLITEN: a preliminary demonstration of sensing strategy for the'energy literacy through an intelligent home energy advisor' (ENLITEN) project, in: Proc. e-Energy '13, ACM, 2013, pp. 279–280.

Figure 1: Experimental Settings – The placement of sensors at each home: (a) Home A – 3 Bedrooms, 3 Floors, (b) Home B – 3 Bedrooms, 2 Floors, (c) Home C – 4 Bedrooms, 2 Floors and (d) Home D – 4 Bedrooms, 3 Floors in Table 1

Figure 2: Sensor box *in situ.*, showing PIR, temperature, CO_2, light, sound and humidity sensors.

Category	**Example(s)**	**MTUS code** [19]
Wash	Bath or shower	Selfcare
Windows	Opening or closing windows and external doors for extended periods of time	–
Eat/Drink	Eating meals, e.g. breakfast	Eatdrink
Food preparation/cooking	Preparing meals	Foodprep
Wash dishes	Using a dishwasher	Foodprep
Cleaning	Vacuuming	Cleanetc
Laundry	Using a washing machine, tumble dryer or iron	Cleanetc
Sport/exercise	Using a treadmill	Sportex
Receive friends	Hosting a party	Leisure
Music listening	Listing to radio or stereo	TVradio
Watch TV	Watching TV, DVD or web-streamed content	TVradio
Play computer games	Using a games console	Compgame
Use computer	Using PC or laptop for work	Compint
Unoccupied	Empty home with no activity	–

Table 2: Energy events logged by study participants, with categories, example events and corresponding MTUS codes.

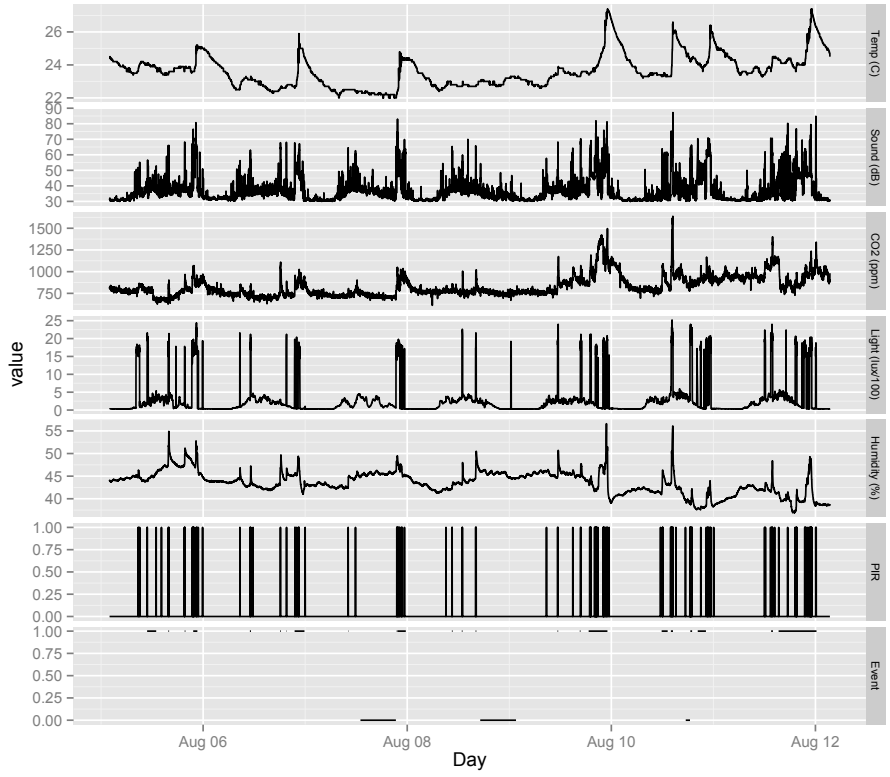
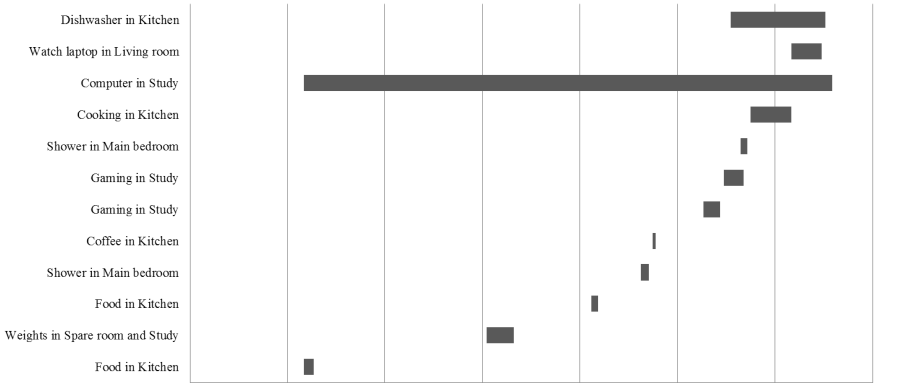


Figure 3: Time series for Home 1's kitchen over the week-long study. 'Event' is encoded as one of three states: event (1), non-event (0) and no record (NA).



	Food in Kitchen	Weights in Spare room and Study	Food in Kitchen	Shower in Main bedroom	Coffee in Kitchen	Gaming in Study	Gaming in Study	Shower in Main bedroom	Cooking in Kitchen	Computer in Study	Watch laptop in Living room	Dishwasher in Kitchen
Start	10:00	14:30	17:05	18:18	18:35	19:50	20:20	20:45	21:00	10:00	22:00	20:30
Duration	0:15	0:40	0:10	0:12	0:05	0:25	0:30	0:10	1:00	13:00	0:45	2:20

Figure 4: Example - Home 1's participant-recorded energy event in a day (Aug 06, 2013)

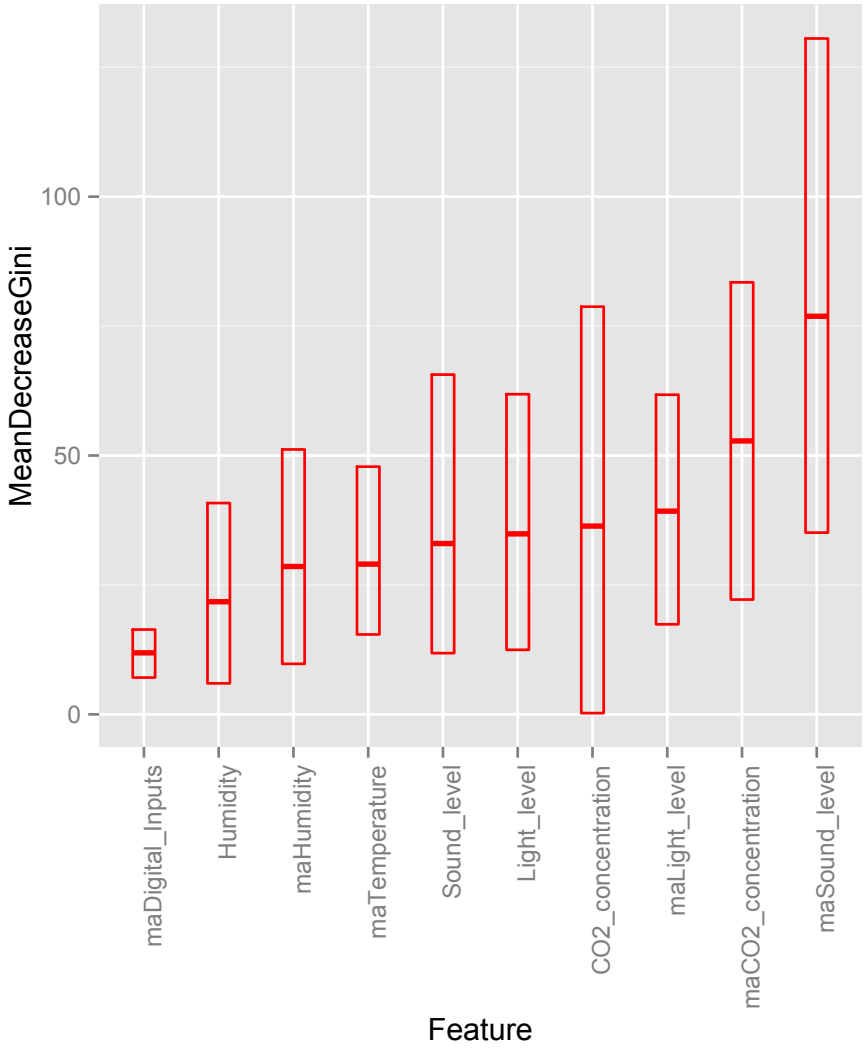


Figure 5: The top ten sensor features using the mean Gini impurity decrease from the random forest process; the larger the better. ma = moving average, .95 CIs shown (non-parametric bootstrap; 1000 replicates).

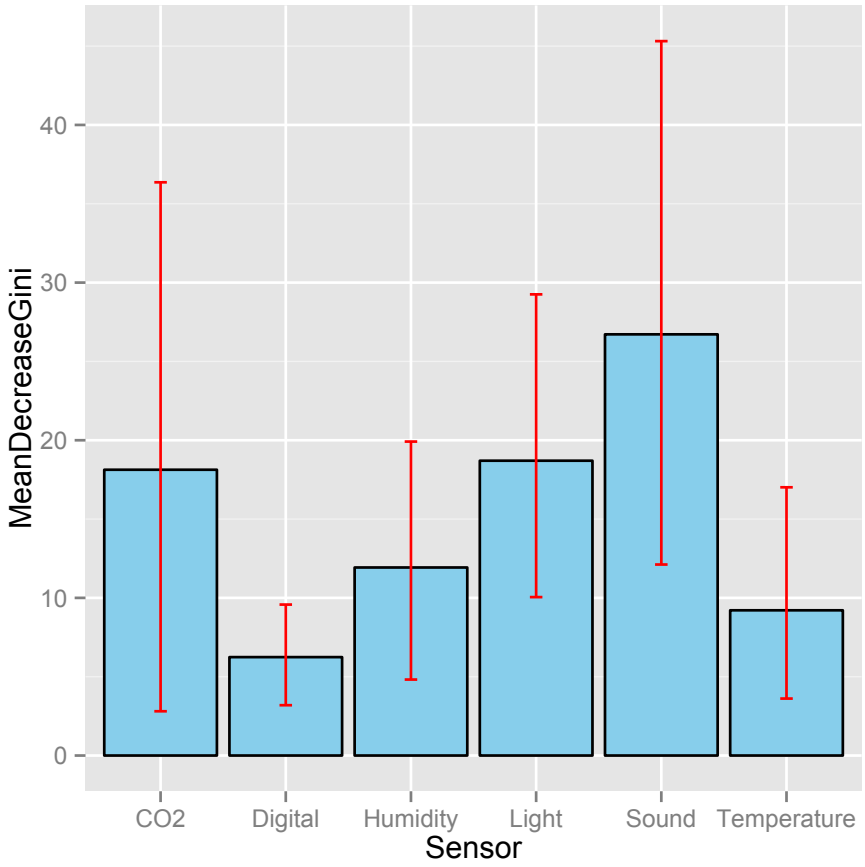


Figure 6: Mean Gini impurity decrease over all features for each sensor. .95 CIs shown (non-parametric bootstrap; 1000 replicates).

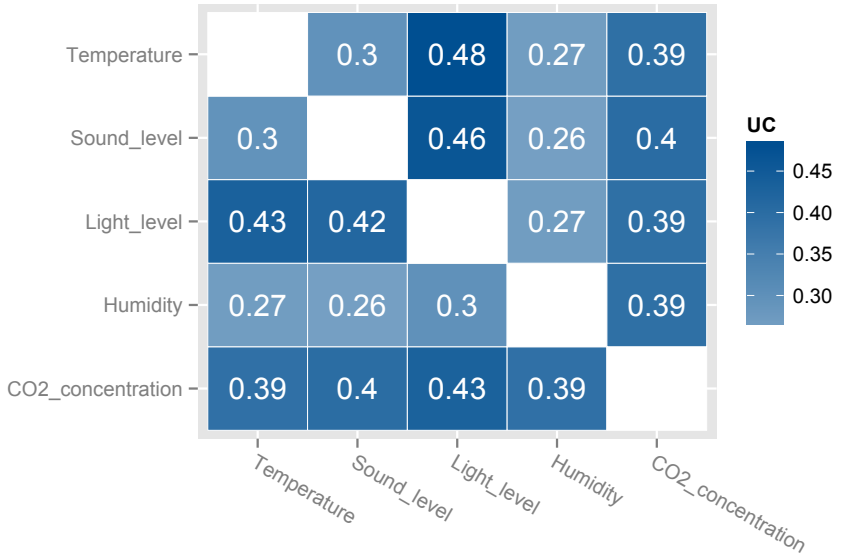


Figure 7: The uncertainly coefficient matrix for co-present sensors. This shows an estimated proportion of bits that can be predicted about sensor j (columns) from sensor i (rows). Note, this function is not necessarily symmetric, and we have omitted the on-diagonal elements and PIR sensor for scale clarity (all < 0.1).

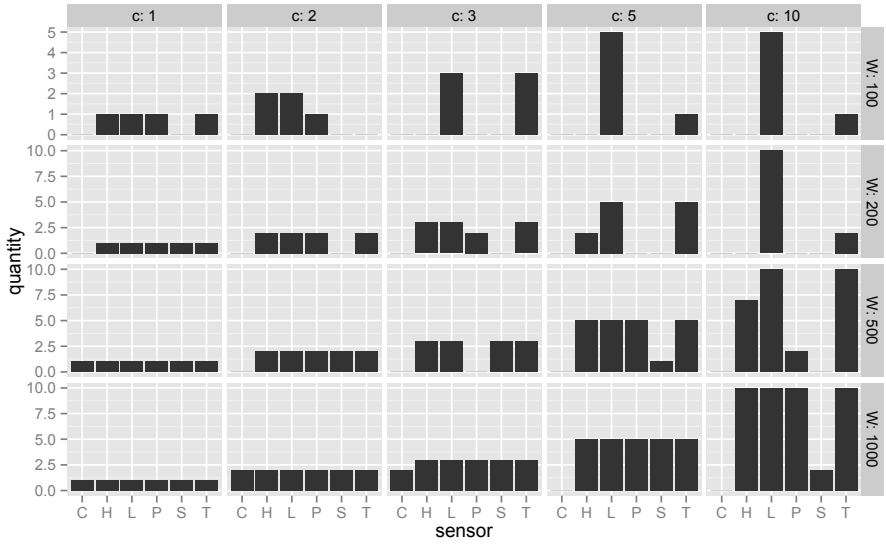


Figure 8: Example sensor sets as output by the BKP algorithm using cost and value data described in the text. C is CO_2, H is humidity, L is light, P is PIR, S is sound and T is temperature.

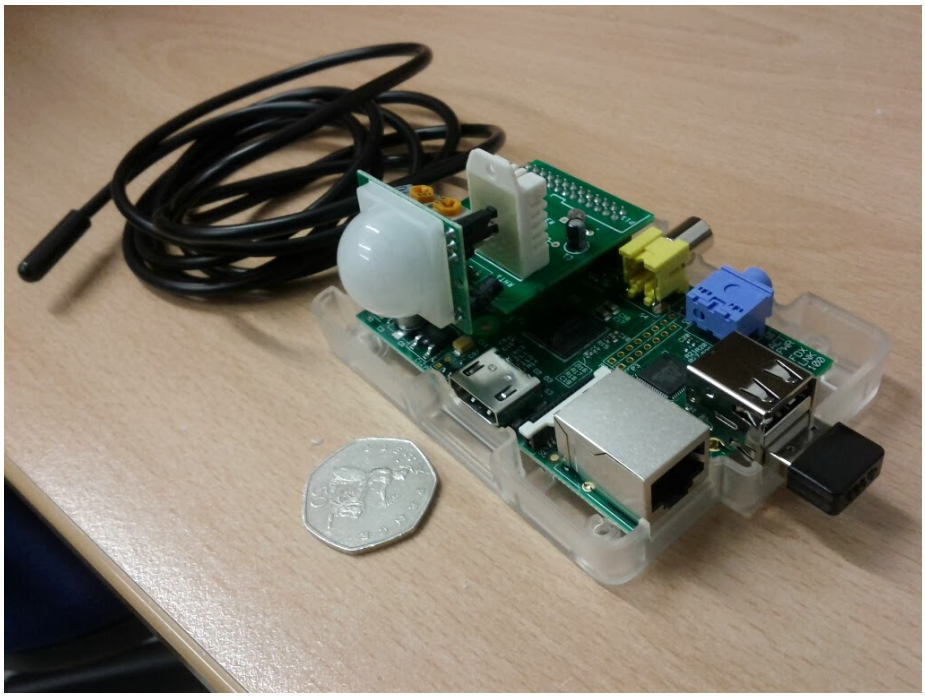


Figure 9: A Raspberry Pi computer showing the cutaway sensor board that is currently being deployed on the ENLITEN project.