The Analysis and Interpretation of Half Hourly Utility Data in UK Buildings

Andrew Wright, Neil Brown
IESD, De Montfort University, Leicester, UK

Abstract
An increasing proportion of utility metering in the non-domestic building sector is at high frequency, typically half hourly in the UK. Originally, this was mainly for electricity in relation to market trading and variable rate tariffs, but it has now spread to gas and water with more emphasis on monitoring and management of utility usage. High frequency meter reads have several potential uses: provision of accurate and timely data; access to trading and variable rate tariffs (normally electricity only); identification of system inefficiencies and failures; quantifying net export from on-site renewables; and long term benchmarking and monitoring. Leicester Energy Agency, in the English Midlands, has over the last few years installed half hourly gas, electricity and water metering in most of its buildings, and now have over 200 buildings monitored in this way.

They have kindly provided a database of readings to the authors at De Montfort University, to use in a major research project called Carbon Reduction in Buildings (CaRB). This provides us with 1349 metering channels of data, representing 598 metering sites, including sub-sites and separately metered outbuildings, mainly for the years 2001 – 2006.

This paper describes some analysis of this rich dataset. Firstly, raw data contains errors and gaps, and various techniques for dealing with these are described, such as removing spurious 'spikes' due to missed reads, and interpolating missing data over short intervals. Four types of building/system failures are then identified which can be inferred from meter reading analysis, and are likely to cause wasteful consumption. Combining data from different utilities is shown to be useful; for example, water consumption serves as a good proxy for occupancy in many premises, and as such can be used to identify when gas heating is on while the premise is unoccupied. Variability between premises of a similar function, and what constitutes a 'typical' profile for a given function (useful for benchmarking), are considered. The use of metrics for normalisation of consumption (floor area, annual consumption) is discussed. The paper includes practical suggestions and many graphical examples.

Introduction
The Leicester Energy Agency (LEA), run as a partnership between Leicester City Council and De Montfort University in the English Midlands, have half hourly metering systems installed in around 270 premises, mostly in Leicester, and mostly owned by Leicester City Council. There are 1349 metering channels of data, representing 598 metering sites, although these include sub-sites and separately metered outbuildings. These channels record half hourly consumption of gas (in m^3), water (in L) and electricity (in kWh), in most cases for the period 2001 – 2007. External temperature data is also recorded from a site in Leicester. It is worth noting the distinction between 'buildings' and 'premises'. Building refers to the physical entity, while premises (confusingly, always plural in this context) refers to the legal entity which may comprise several buildings, or part of a building, at the same site. In cities particularly, many premises (e.g. offices, shops) are in larger buildings, but certain types of premise such as schools occupy several buildings. It is worth noting that European Directives do not make the important distinction but just refer to buildings. Most of the data described was for a single [set of] premises, but sometimes there was sub-metering on a large site.

Water consumption is relevant for several reasons. The water industry is energy intensive and consumes about 2% of total energy used in the UK [1], mainly for pumping and treatment. According to the UK water industry, the average energy consumption is 586 kWh/Ml (mega litre) for water supply and 634 kWh/Ml for sewage treatment – or 1220 kWh/ML total, about 0.0012 kWh per litre water used [2]. Of more immediate use in analysis is the fact that water consumption can be used as an informal indicator of occupancy. And finally, water may become a scarce resource if droughts become more frequent in a changing climate.

The premises include typical local government premises such as schools, offices, libraries, but also commercial premises such as industrial units, and shops. As part of the Carbon Reduction in Buildings (CaRB) Project [3], metering and other data from these systems have been loaded into a database, with a view to producing improved understanding and models of energy use in the UK building stock. The database is available through a database server running MySQL [4], on a RedHat Linux operating system. It contains around 20 million data points.

Premise types

The types of premises connected to the metering system are shown in Table 1. Floor area data, where available is internal. The Primary Classification (Pclass) codes shown in this table refer to building classifications developed by Harry Bruhns at UCL [5]. Each site in the database has been categorised thus according to its corresponding usage code. At this point in time there are not detailed data on age, construction, HVAC equipment etc. but surveys are planned to fill in these gaps.

Table 1: Distribution of premise types.

Pclass code	count	mean floor area	usage type
CO	3	3994	Offices
CO2	19	2842	Local government office
CR141	1	2000	Indoor markets
HL11	3	3288	Museum or art gallery
HL12	16	565	Library
HL30	3	2497	Leisure centre
HL311	1	2596	Sports centre with pool
HL312	2	1465	Sports centre pool LA
HL314	1	2000	Sports centre - no pool
IF	1	2541	Manufacture
IF1	33	121	Workshop
IT22	1	186	Bus depot
SE11	1	480	Nursery or kindergarten
SE2	41	2164	State primary school
SE31	15	9540	State secondary school
SE32	2	3092	Special school
SQ	1	1597	Community facilities
SQ10	15	1719	Community centre
SQ11	1	724	Hall

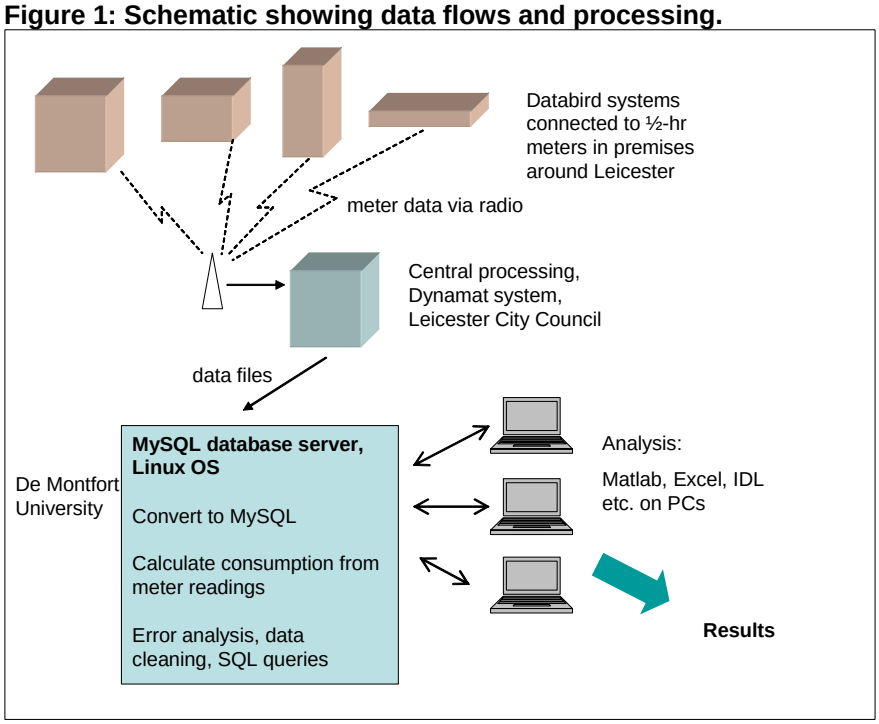
Data handling and cleaning

This section describes the general approach to handling, cleaning and analysing the data. The data collection and analysis process is shown schematically in Figure 1. Data from meters are logged and transmitted via a network of seven radio stations around Leicester to a central system run by LEA. Periodically, the data are transferred to IESD at De Montfort University in text format, and converted to MySQL database format. Data are held on a secure system for access from the local network. Data are cleaned and some basic analysis can be done in SQL. More complex analysis is done with other software at individual computers.

General approach and software

LEA, and most commercial organisations, use specialist software provided with the metering system to analyse the data. This is well suited to day-to-day operations such as triggering alarms when unusual consumption occurs and general monitoring, but less suitable for scientific analysis – also the software is usually expensive to buy.

Figure 1: Schematic showing data flows and processing.



For relatively small amounts of data, or large datasets with a simple structure and mainly numerical data, it is feasible to use a range of software tools such as Microsoft Excel, Mathworks Matlab, various statistics packages, Unix scripting combined with tools such as AWK and GREP, etc. But for a large complex dataset with many text fields, a relational database is likely to be much more efficient. Virtually all databases have Standard Query Language [4] (SQL) capability which is powerful for querying across different tables of data and dealing with text fields. MySQL was chosen because it is a low cost, powerful, open source system which runs across various platforms – the server was running Linux but most of the analysis was done from Microsoft Windows.

Data were stored consecutively, as meter readings, with the site identifier, date-time, utility type, conversion factors, etc. given in other fields, and related tables for premises etc. So one day of readings for say electricity for one meter was represented as 48 rows of data; one year by 1760 rows of data. An important data handling task was the addition of a numerical indexing key for each metering data entry and metering channel, such that energy usage may be found in a reasonable amount of time – this reduced some queries from hours to minutes.

Databases have date and time functions, so it is possible for example to get monthly consumption of all sites, or total energy use for weekdays in a particular month and set of sites, in a one-line query. However, for detailed numerical analysis, such as removing errors or analysing daily profiles, SQL is less suitable. In particular there is no notion of time series in databases; there are just data with timestamps, and the records could be in any order (though order will affect efficiency).

For this work Mathworks Matlab software was used for data cleaning and analysis, though others could be used – another project used Research Systems Inc. IDL software on a subset of the data. Matlab has many powerful data analysis features and is efficient for handling large arrays. It can also be linked directly to the database.

Data quality problems

In a database of this size, with data coming from field equipment over a wireless system, and with considerable human input of information on the sites etc., there are inevitably data quality problems. Some of these could have been avoided by a better original data design, but this is easy to say with hindsight. The most significant of these were:

- Periods of zero readings – it is not always clear whether these represent genuine zero consumption (for example not heating during summer; premises vacant) or missing data. Electricity consumption rarely goes to zero, but water and gas often do. The most reliable way to identify missing data is to compare with similar days and with consumption for other utilities.
- Inconsistent values in text fields – for example names of premises
- Missed readings, typically for one or two half hours; when converted to consumption every half hour a missed half hour appears as zero consumption followed by double the normal consumption.
- Microprocessor errors which can result in spurious numbers in the readings.
- Conversions – the information fed back by the radio system was the number of meter pulses counted, but what a pulse represented varied considerably due to meters of different ages and varying size of premises. Gas was measured in various fractions of cubic feet and cubic meters with scaling factors in powers of 10 ranging from 0.01 to 100. Electricity was measured in kWh, kVA and kVAr (reactive power) depending on the site, with a large number of scaling factors such as 1, 0.1, 0.2, 0.4, 0.5, 0.15 etc. Water was measured in litres with factors of 5 or 0.5, or powers of 10. In total there were 53 different scaling factors – careful checking was needed to ensure that the wrong scaling factor was not used.

The greatest problem for analysis is missing data. A balanced approach was taken, interpolating short periods of missing data and including the interpolated data in the analysis, but excluding longer periods of missing data from the analysis. This approach avoids 'throwing away' large amounts of essentially reliable data with just a few errors, while retaining statistical reliability. (An alternative would be to 'fill in' large gaps with data duplicated from similar periods, e.g. winter weekdays. This would simplify analysis but bias the results towards the duplicated periods since these would be over-represented.)

Text field problems

Table 2 shows examples of different descriptions of the same premises (there were many others). In retrospect, it would have been better to have a separate table listing all the premises just once, then link this to the tables of data. The Blackbird Road Depot is for training gas technicians, with the potential for confusing the meter type with the premise description. This sort of problem is not unusual.

Table 2: Examples of ambiguous naming for meters at two sites.

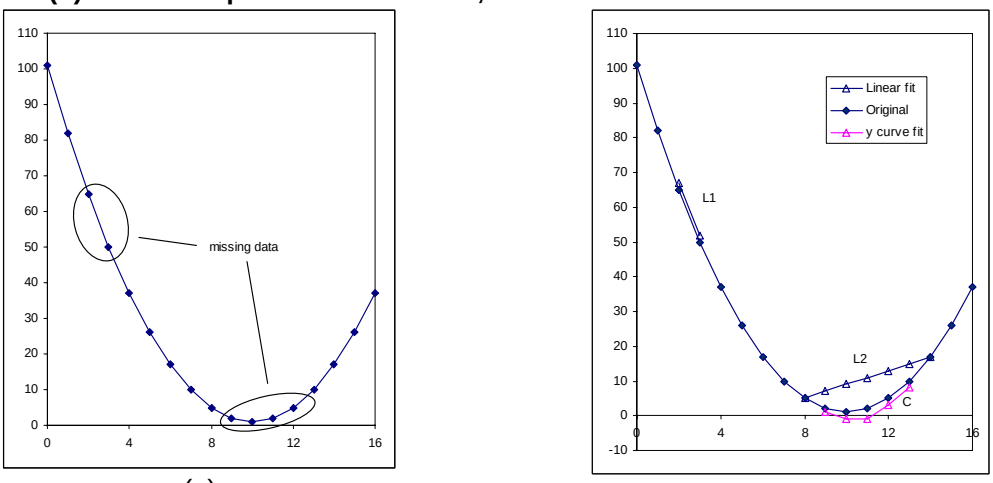
```
gas - Abbey Primary School:
--------------------------------
1 --- Abbey Primary Sch Main Gas
2 --- Abbey Prim Ross Walk Gas
3 --- Abbey Prim Community Gas

gas - Blackbird Road Depot:
--------------------------------
1 --- Bl'bird Rd Gas Training Gas
2 --- B'bird Gas Training Water

electricity - Blackbird Road Depot:
--------------------------------
1 --- Blackbird Road Depot Elec

water - Blackbird Road Depot:
--------------------------------
1 --- Blackbird Road Depot Water
2 --- B'bird Gas Training Water
```

Data interpolation

There are various ways of interpolating missing data, but is not simple because many of the curves are 'spiky' with discontinuities in the slope (first derivative). Figure 1 (a) shows a parabolic curve, for illustration, with areas where actual data might be missing ringed. Linear interpolation works well where the second derivative of the curve is close to zero (line is close to straight) as shown in Figure 1 (b) at L1; but linear interpolation at L2 as the curve approaches a minimum is very different from the actual curve. In general, linear interpolation only works well when the second derivative of the curve is close to zero (slope of line is about the same on either side). A curve fit, such as a cubic spline, works better at following the curve, as shown at C in (b). However, it can occasionally result in impossible, for example negative, values, as shown here, and can produce false maxima or minima. A linear fit is

guaranteed to produce only values between the two end points. Another problem with cubic splines is that it can give odd results if data are missing at the start or end; such values were removed for the analysis. For most purposes, cubic splines were used as these gave a much better fit to the data; 'impossible' negative values were replaced by zeros, although negative values were rarely generated in practice.

Figure 2: (a) Original data showing where gaps are and what correct values should be within ellipses, and (b) linear interpolations L1 and L2, and curve fit C.

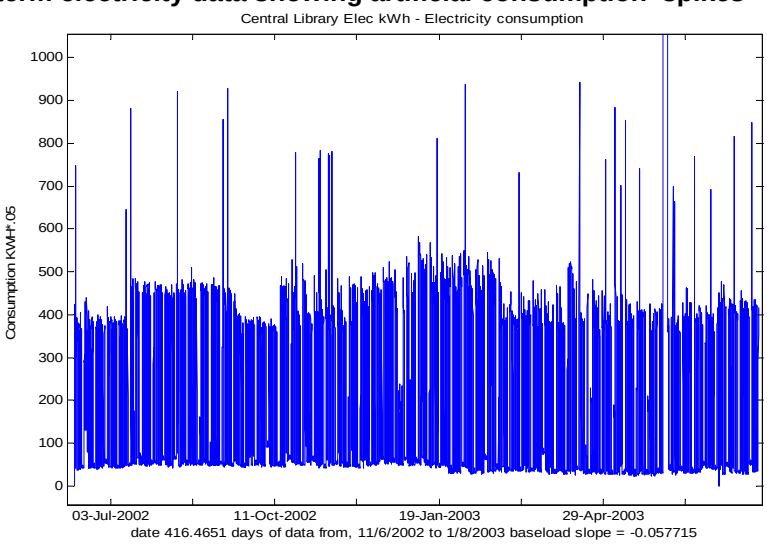


(a) (b)

Data spikes due to missed readings

Usually, radio based collection is reliable, although dropouts may occur. A dropout is when a data unit is logging metering data locally, but is not able to send meter pulses by radio back to the central radio receiver. As a result, a signal delayed for an hour, for example, will contain the number of meter pulses logged during an hour, instead of the normal half hour. So when plotted there will be a data point where radio transmission has resumed, resembling an integrated 'spike' of high consumption, and the timestamp will jump to the next transmission time. Figure 3 shows an example from just over a years worth of data from one of the premises in the Leicester database, where data 'spikes' are clearly visible.

Figure 3- Long term electricity data showing artificial consumption 'spikes'



Various attempts have been made by researchers in the field to find a rapid yet reliable way of cleaning this data without manual intervention. One approach for error detection is if the metered reading is away from the norm by a set number of standard deviations, a spike is assumed, and this

part of the data is trimmed. Following this, some favour averaging the readings which would have been obtained during a dropout period, others favour a linear interpolation. These methods are fairly crude but robust, although for the majority of cases a more sophisticated approach yields corrected results which appear to follow much more closely the typical consumption profiles. Of course there is no better method than making sure that data is available for every available time step in the first instance.

The procedure adopted was as follows:

1. Since a time error will occur at the same time as an integrated spike, we can find all time errors by taking the first difference of the time axis and putting this into an array (X1).
2. Then 1 is subtracted from all array elements X1.
3. Zero (normal) values are left as zero, positive (missed data) values are set to 1. This is a very rapid operation in software.
4. A logical NOT is performed, e.g. $1 \rightarrow 0$, $0 \rightarrow 1$.
5. The metering data (Y1) is gated by this array, i.e. where a 0 is present, data is not allowed through. This neatly trims off data spikes.
6. A new time axis (X2) is generated as an equally spaced vector, with start and stop times equalling the original sampled data in (X1, Y1)
7. A cubic spine is used to resample the data to produce a new array of metered data.

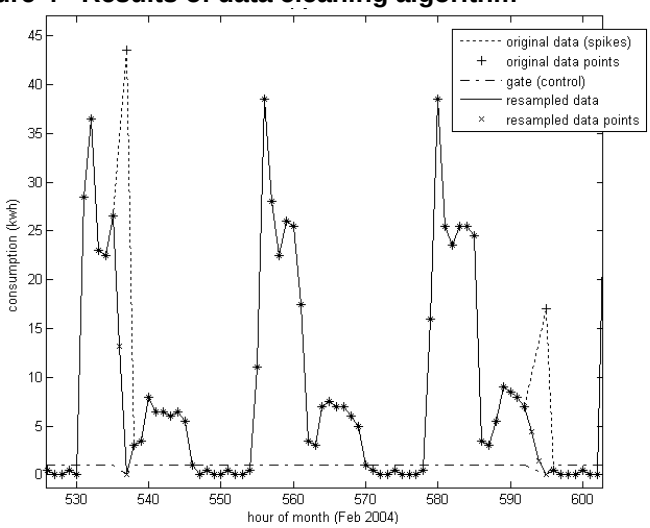
Table 3 shows how this process works in practice for a simple example of a time array containing an error shows the result of this operation and illustrates steps 5, 6 and 7 above.

Figure 4 time steps with missing data are shown by the dotted lines. The gate array, used to trim data spikes, is shown as a dotted and dashed line, and original data points as a +. As can be seen , the effect of resampling (shown by a X for each data point), closely follows the original profile data. Where Y1 has been trimmed due to missing time steps, the interpolated data can be seen to be providing a reasonable approximation to the typical profile. An item for future research is to carry out a large statistical analysis to further validate the robustness of this method.

Table 3 - Data cleaning algorithm

Step	Operation	Sample of array
	(original time axis, gap starts at 4)	1 2 3 4 7 8 9 10 11
1	Differences	1 1 1 3 1 1 1 1
2	Differences -1	0 0 0 3 0 0 0 0
3	Sign of differences	0 0 0 1 0 0 0 0
4	NOT of previous result = gating array	1 1 1 0 1 1 1 1

Figure 4 - Results of data cleaning algorithm



Microprocessor errors

Other errors which may occur in the metering system are usually due to problems with the microprocessor based transmission equipment. A frequent error is incorrect sequencing of data in the raw dataset, such that data (and corresponding time steps) are not always stored chronologically. This is not a problem using the setup at IESD since the database server automatically sorts metering data by corresponding time stamps. It must be remembered however that simple software working on the raw dataset, and possibly some other database engines, will produce profiles with a few steps backwards in time. With regard to all of these errors, a manual check is carried out when plotting data in small batches, but most faulty data can be caught during interrogation of the database with SQL through error trapping.

Results

Building energy failure modes

One early result of analysis is that certain key Failure Modes for building energy consumption are easily identifiable. These modes are detectable with little prior knowledge of the building, and do not require weather or floor area data. The principal Failure Modes identified were:

1. **Heating (or cooling) out of season.** Winter cooling or summer heating. In an extreme example, where narrow dead bands are used in building controls, both heating and cooling may occur at once, particularly during spring and autumn changeovers.
2. **Heating when building unoccupied.** Particularly on Sundays, we have seen evidence of unoccupied premises being heated. In these cases, simply addressing this issue could save around 14% of gas consumption.
3. **Baseloads.** For electricity, this can mean excessive numbers of appliances not switched off, but for gas consumption will usually mean faulty building controls. For the purposes of this analysis, we would usually define a gas profile as unchanging by day or week as a gas baseload, since it represents an 'energy leak'. Even in some sites which are staffed on a 3 shift basis, such as care homes, some periodicity should usually be present in half hourly profiles.
4. **Excessive consumption (continuous).** Typically if no periodic profile is followed, a system may be permanently running – usually gas heating. Two possible causes would be faulty building controls or faulty boiler instrumentation. If a boiler is not modulating but appears not to run at night, then a profile more akin to a square wave will represent daily gas consumption. This means that a boiler would be running at close to 100% whenever switched on, and is controlled by a functioning timer, although other controls would be malfunctioning.

Some examples are now given for these modes. Curves are normalised to consumption per 1000 kWh annual consumption. Data are from actual weeks in summer and winter where there was no missing data. Electrical loads with a clear diurnal pattern are a fairly reliable indicator of occupancy, or at least when lights are on.

Figure 5: Energy plots for a library; electricity, top (red summer, black [heavier line] winter) and bottom, gas summer (no data for winter), showing heating out of season and overnight (Modes 1 and 2); the x-axis 'Hour from Sun 0:00' means hours from midnight Sunday.

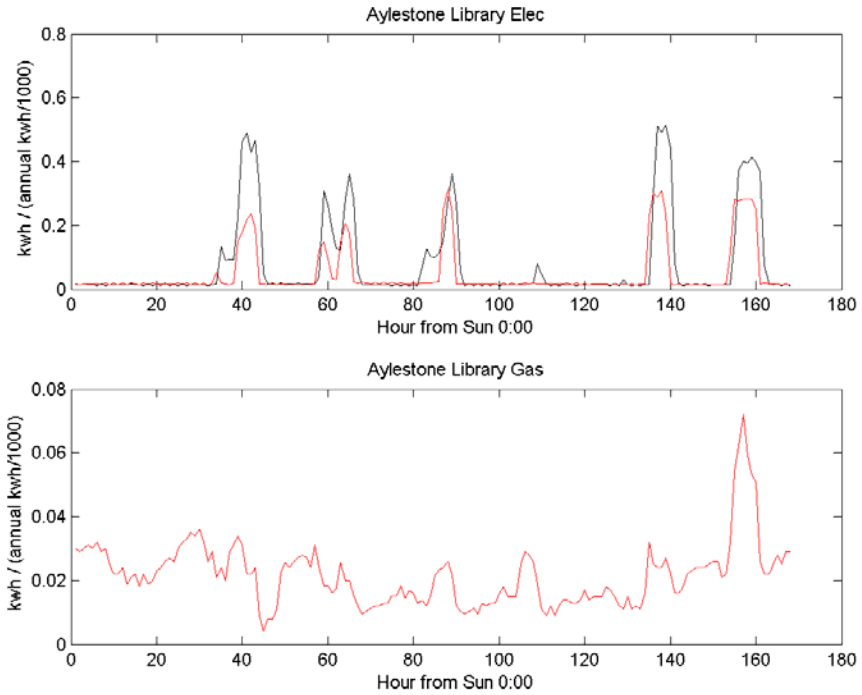


Figure 6: Energy plots for an office block; electricity, top (red summer, black winter) and bottom, gas (red summer, black winter), showing winter heating outside occupancy (Mode 2).

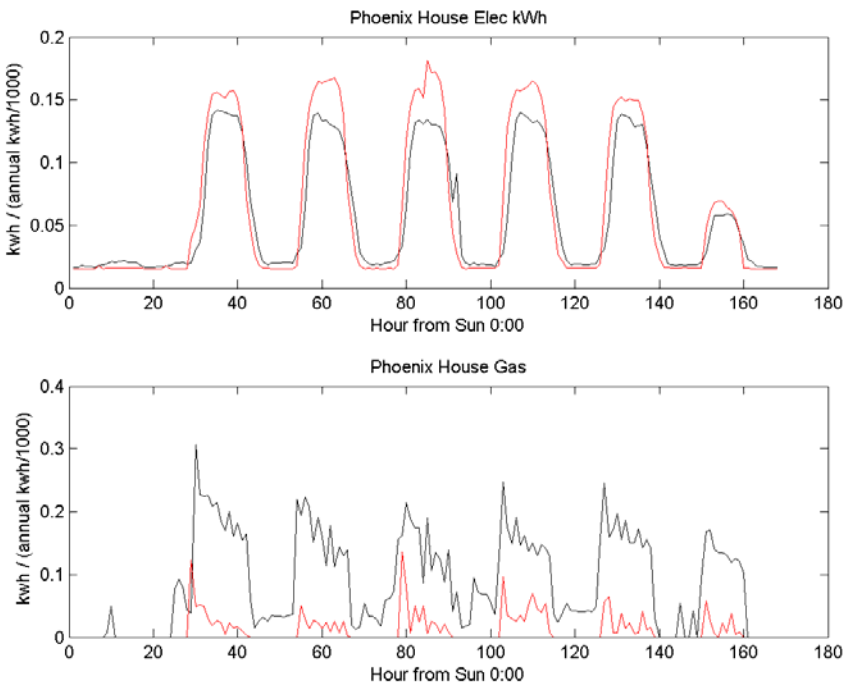


Figure 7: Energy plots for an office block; electricity, top (red summer, black winter) and bottom, gas (red summer, black winter), showing winter heating outside occupancy (Mode 2), and high baseload for electricity in both seasons (Mode 3).

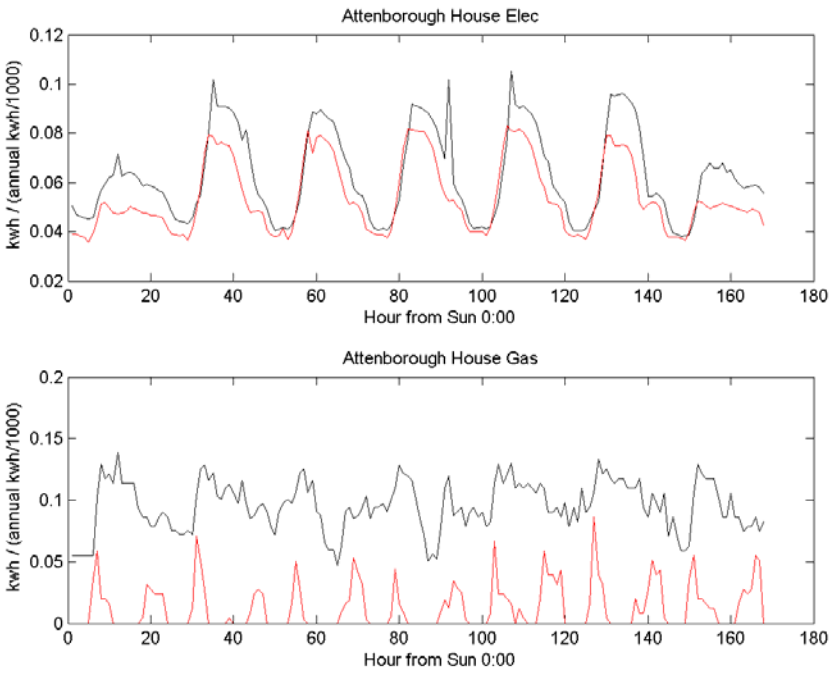
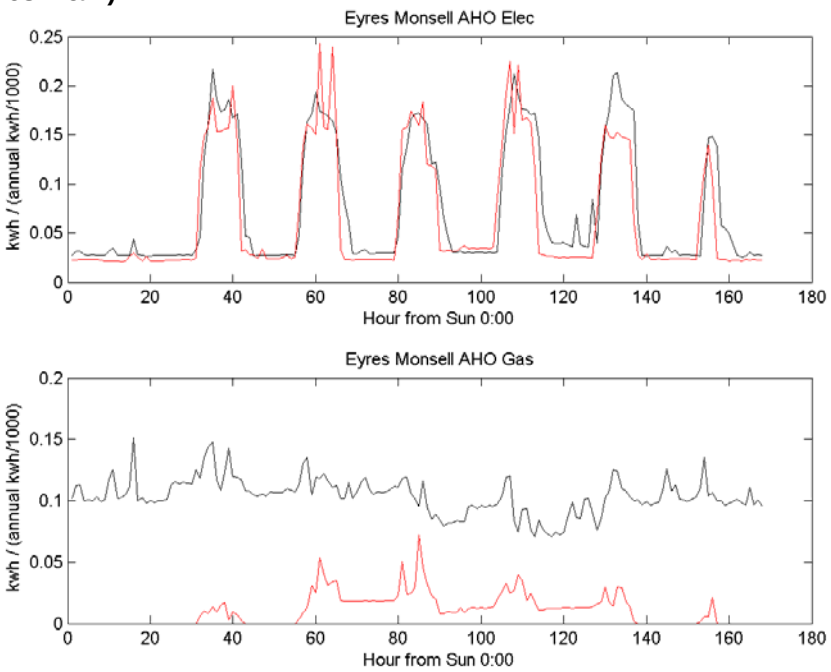


Figure 8: Energy plots for an office block; electricity, top (red summer, black winter) and bottom, gas (red summer, black winter), lack of control of summer and summer heating (Modes 2 & 4).



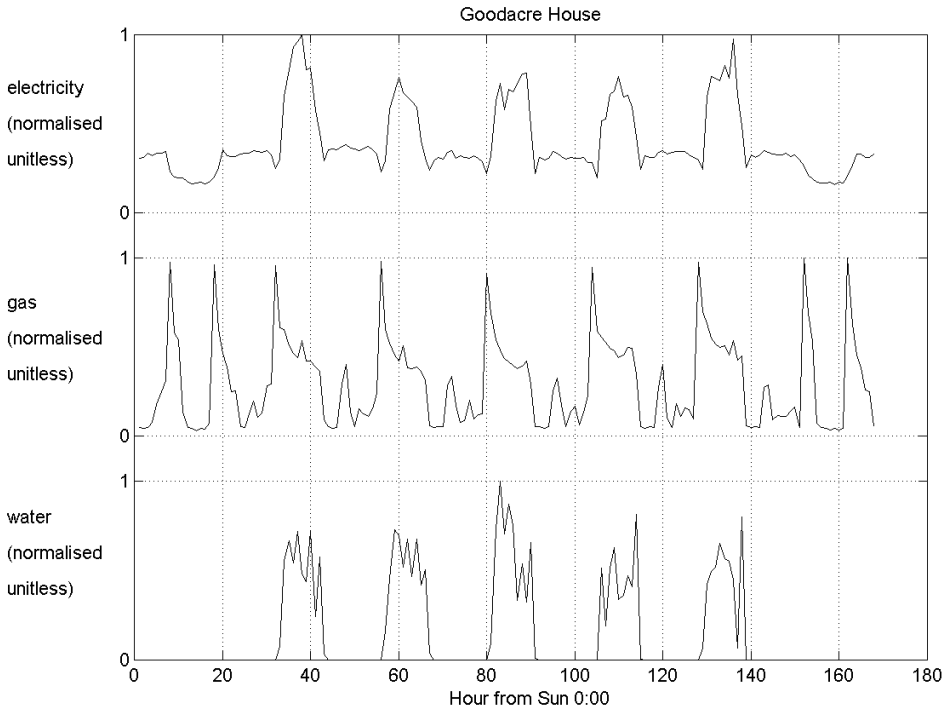
Use of water consumption data

Monitoring and reduction of water consumption through the half hourly metering system is carried out primarily to reduce costs and prevent leaks. Interestingly, some of the largest savings for LEA have come from the identification of leaks – as these often occur underground they can go undetected for a long time.

However, water consumption profiles are also useful indicators of occupancy. A peak of water consumption will occur at times of maximum occupancy - this is routinely used by the LEA to look for situations where empty premises are heated.

A baseload for water consumption usually means a leak - this occurs fairly frequently, sometimes due to cast iron Victorian water mains and heavy traffic and building works. Frequently however large losses can occur from the accumulation of seemingly minor losses. Examples would include dripping taps (to 80 litres per day), dribbling taps (to 300 litres per day) and faulty cisterns (to typically 200 litres per day). From Figure 9 it is clear that the premise is unoccupied at weekends and overnight (and there are no significant leaks), but there is a very high electricity baseload and gas heating running 7 days a week.

Figure 9: Office block electricity, gas and water consumption for one week in February 2004.

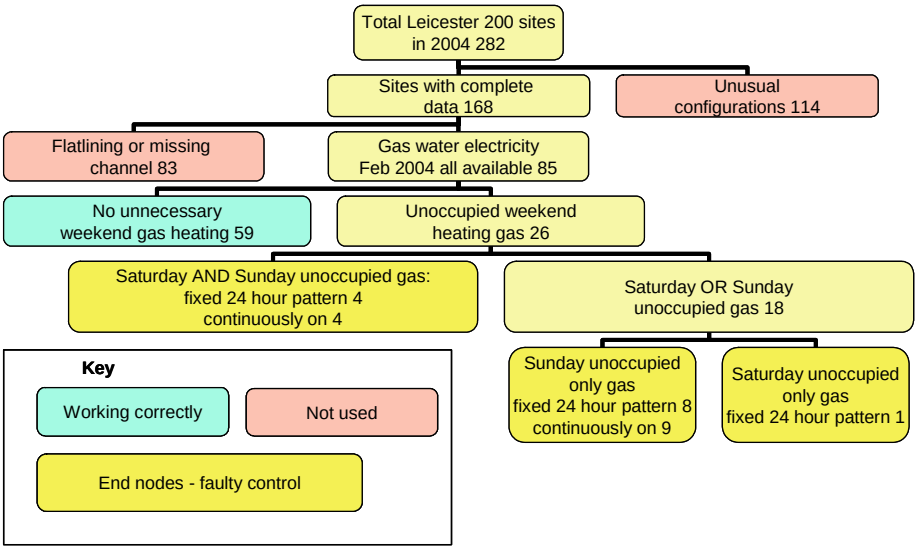


Analysis of control failures across a large sample

A large proportion of the premises were analysed to see how prevalent control failures in gas heating were. The successive breakdowns of the sample is shown in

. Out of 282 sites, there was complete data for 168 sites; of these about half, 85, had gas, water and electricity data for February 2004. Analysis of these 85 showed 26 (31%) had heating during unoccupied periods at weekends. These can be divided into 8 (31%) with heating on both weekend days, and 18 (69%) with one or the other days. Of these, 17 had heating on during Sunday only, and one on Saturday only, varying between 24 hour heating and timed heating. Of all the premises with unoccupied weekend heating, 46% had timed heating and 54% had heating on all the time.

Figure 10: Analysis of gas heating control failures across large sample, broken down from top to bottom.



Out of the sample of 85 premises with 5 day occupancy analysed in detail, if there was no weekend heating there would be, each week, 85 × 5 = 425 'premise heating days'. But there were actually 85 × 5 + 8 × 2 [Sat & Sun] + 18 × 1 [Sat or Sun] = 459 premise heating days. Very approximately this represents 8% more heating energy usage than needed, if one assumes that heating energy is proportional to the number of days heated per week. However this ignores additional savings from correct time control over 24 hours found to be absent in 13 (half) of the premises with unnecessary weekend heating, and unnecessary heating in warm weather. Therefore it seems very likely that more than 10% of gas energy usage could be saved in this large sample simply from correct time control – before even considering more efficient boilers, better thermostatic control, improved fabric etc.

Typical consumption patterns for different premise types

For utility planning or testing the feasibility of community energy systems, it is very useful to know the likely profiles of existing or new buildings. The authors are not aware of any generally published data on such profiles for the non-domestic stock. The electricity industry in the UK uses profiles based on load factor (the ratio of average to peak load), which is only one very general characteristic of the load pattern and not directly related to the activity of the customer.

Normalisation

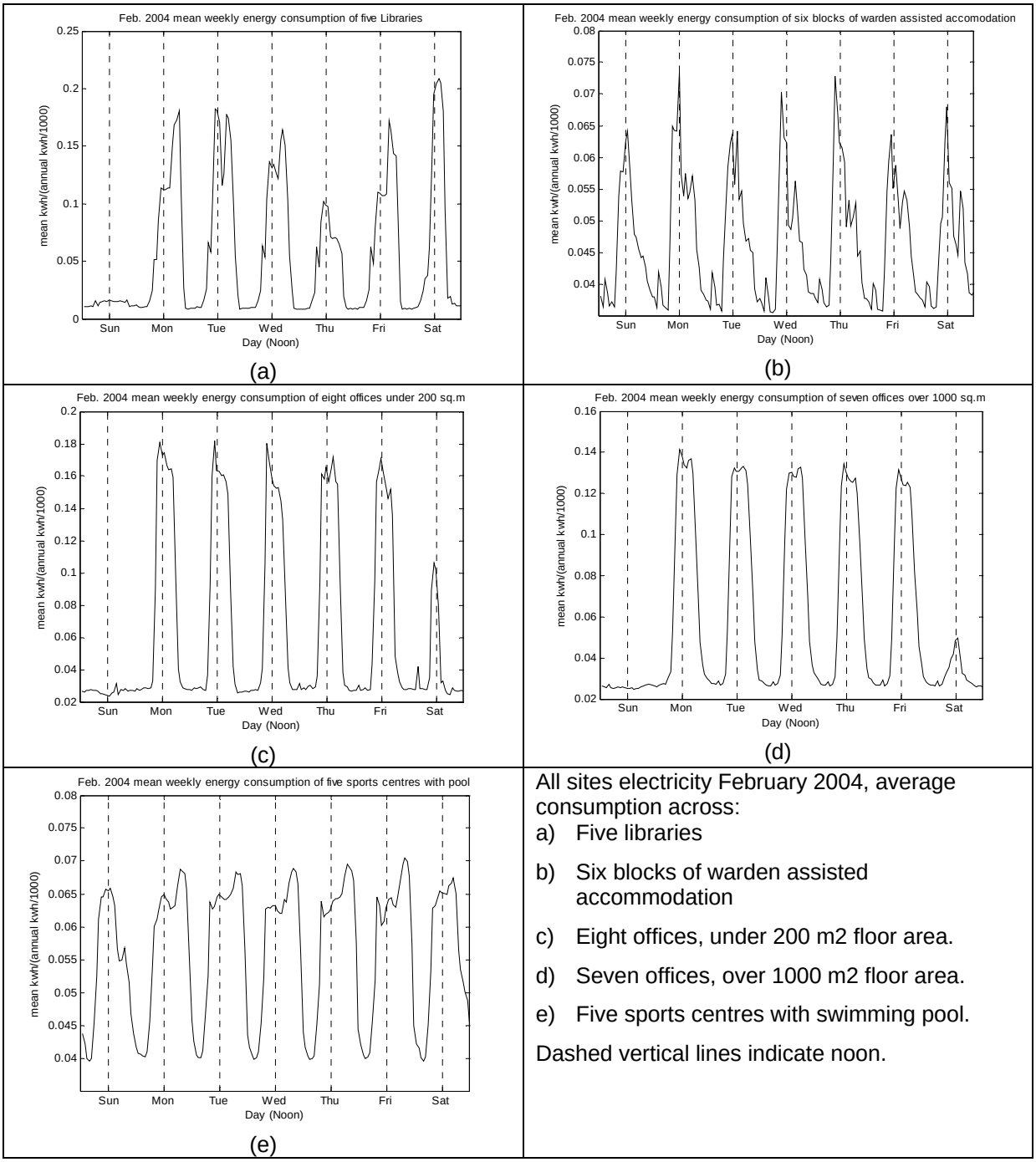
It is common practice to normalise building energy consumption by floor area. However, when comparing load patterns it makes sense to normalise by annual consumption, for example to replace kW with kW per 1000 kWh of annual consumption. This approach gives equal weight to the shape of each pattern. Then for another building where the annual consumption is known, the average of the normalised curves can be scaled back to give a typical consumption pattern in kW, such that if this is integrated over a year, the actual total consumption would result.

Consumption patterns

As shown previously, consumptions were normalised by annual consumption for each premise. These could then be simply averaged across premises of similar type to produce a representative profile, without being dominated by the premises with larger usage. As they are averaged from a fairly small number of premises, they cannot be taken as statistically reliable, but do illustrate the proof of concept. To get the actual profile for a given premise, multiply the profile according to the scaling factor given on the y-axis. Figure 11 shows average consumption patterns for five types of premises. Figure 11 (a) for libraries shows there is no Sunday opening, possibly a shorter day on Thursdays, and peak load on Saturday – probably the busiest day. There is a less clear pattern in (b) but highest use is around noon, probably for cooking, with a small evening peak which may be related to cooking also, and the lowest baseload of all the types. Small (c) and large (d) offices show a very similar, classic rectangular pattern, with a smaller Saturday peak probably due to a shorter opening period.

Sports centres (e) show a high base load, probably for pool systems, and higher use in the afternoons and around midday on Sunday which would correspond with typical usage patterns. The ‘spiky’ nature of some plots suggests a larger number of premises is needed to give a reliable typical shape, particularly where patterns are likely to vary a lot between premises.

Figure 11: Average, normalised electricity profiles in one week for different premise types, February 2004.



All sites electricity February 2004, average consumption across:

a) Five libraries

b) Six blocks of warden assisted accommodation

c) Eight offices, under 200 m2 floor area.

d) Seven offices, over 1000 m2 floor area.

e) Five sports centres with swimming pool.

Dashed vertical lines indicate noon.

Conclusions

The power of using a relational database, combined with data analysis software (Matlab), to analyse large quantities of half-hourly metering data has been demonstrated for premises in Leicester. Complexities of data quality and structure, utility conversion factors and interpretation for this sort of dataset are significant. Cubic splines were found to be a good way of interpolating small periods of missing data. However, analysis is best confined to periods with mostly 'good' data to minimise assumptions about other periods. Actual individual weeks in summer and winter, with no missing data, were found to be effective for forensic investigations and visual analysis. Four failure modes were identified with examples, where wasteful consumption occurs as a result of poor time or thermostatic control, unnecessary overnight and weekend use of electricity, or leaks (water). Electricity and water are shown to be useful indicators of occupancy and hence wasteful consumption. Analysis of heating in 85 premises showed around 8% of excess gas use purely due to incorrect heating time clock settings causing weekend heating on unoccupied days. Normalised weekly consumption for premises with similar function can be averaged to give a more 'typical' consumption pattern, useful in energy planning.

Most of the data analysed was from 2004. The LEA has used the metering data to identify control problems, water leaks etc. and most or all of the problems used in this paper for illustration have undoubtedly been corrected (though it was not possible to measure this as we did not have a record of when corrections were made). Indeed, this is the main purpose of having the system in place. However, the premises monitored make up only a tiny percentage of the total non-domestic stock in Leicester. Very few of the other premises of similar size will have half-hourly metering and active energy management, which strongly suggests there are many other cases of wasteful consumption not being identified.

A great deal of further work could be carried out on this data. This includes more statistical analysis of wasteful consumption; development of typical profiles using parametric models related to season, temperature etc. rather than just averaging for one week; disaggregation of electrical loads such as summer cooling load as a function of external temperature; and use of half-hourly data in relation to benchmarking such as the requirements of the Energy Performance in Buildings Directive and the Energy Services Directive. Ultimately, a combination of a good understanding of a building and its loads, half-hourly consumption data, and efficient analysis systems (ideally online) could eliminate a high proportion of wasteful consumption in non-domestic buildings with consequent savings in cost, resources and carbon emissions.

References

[1] *Water UK website*, http://www.water.org.uk/home/policy/climate-change/mitigation [accessed January 2008].

[2] *Towards sustainability 2005-2006 (UK water industry Sustainability Indicators 2005/2006),* Water UK, London, available at http://www.water.org.uk/home/policy/reports/sustainability/indicators-2005-06/towards-sustainability-2005-2006-highlights.pdf [accessed January 2008].

[3] Lomas, K., Oreszczyn,T.,Ward,I.C., Guy,S.C., Shipworth,D. *Carbon Reduction in Buildings (CaRB): A socio-technical, longitudinal study of carbon use in buildings*. 2004 [cited; Available from: http://www.carb.org.uk/.

[4] *MySQL: The complete reference*, Vaswani V, McGraw Hill/Osborne, Emeryville, CA, 2004.

[5] Bruhns H R, Steadman P, Herring H, Moss S, Rickaby P A, *Types, numbers, and floor areas of nondomestic premises in England and Wales, classified by activity*, Environment and Planning B: Planning and Design 27(5) 641 – 665, 2000.