NEW METHOD TO RECONSTRUCT BUILDING ENVIRONMENTAL DATA

Alfonso P. Ramallo-González
Department of Architecture and Civil Engineering, University of Bath, Bath, UK.

ABSTRACT

Access to data from real measurements is in most cases incomplete. In this work, the gaps that were found in data from two real projects were studied. After studying the nature of the gaps and their probability distributions, five different methods for the interpolation of the incomplete data have been evaluated.

None of the methods has been seen to be fail proof. However, the results shows that gaps smaller than six hours are rather easy to fill, but that above this length methods that are more complex need to be used. ARIMA models seem to be one of the best options to fill substantially long gaps.

INTRODUCTION

Assessment of building energy demand or comfort levels using building simulation using energy modelling is becoming a common practise among architects and engineers (Crawley, Hand et al. 2008).

Moreover, the use of these tools, either through direct modelling or through data driven modelling, requires of input data. In the majority of the cases that data comes in the form of time-series.

Artificial time–series synthetically generated are in some cases used as inputs in building models (such as weather files (Eames, Kershaw et al. 2011)); but also data from the real world is used for calibration purposes or data-driven models.

Real sensors in the real world fail. The work presented in this paper shows a study on the gaps that appear in data sets. When one observes data from the real world, two broad groups of reasons for failure could be seen:

- Technical
- Human

Internet connexion being lost, recording problems, hardware malfunctioning and/or software failing could be some of the reasons for gaps in data due to technical problems.

Additionally, human problems are substantial when measuring variables in real dwellings. The normal life of the occupants can interfere with the correct reporting of the sensors. Examples of these are occupants disconnecting the sensors because they want to plug something else, router being turned off, accidents that damage the sensors or just rejection of the sensor.

The rational of this work, is that most building simulators require complete time series to run calibrations or as inputs. In addition, inverse modelling benefit of complete time series too.

In this paper, we have used two data sets that have been obtained by different research teams: (1) data from the Micro CHP Acceleration Plan by the Carbon Trust (CHP from now), and (2) data from the ENLITEN project by the University of Bath.

Firstly, an analysis on the gaps of the two data sets was done. Secondly, we compared several reconstruction methods to evaluate the most optimal method of reconstructing the time-series depending on the data being reconstructed and the length of the gap.

METHODOLOGY

Analysis of the gaps

Data collected from 86 buildings, 30 from the Carbon Trust project and 56 from ENLITEN, together with environmental external data obtained from a server were analysed to understand the nature of the gaps in the series.

Analysis of the data from the CHP project

Three variables were analysed in this data set:

- Internal temperature,
- External temperature,
- Electricity use.

Each one of the time series in the 30 homes was proccessed to obtain the number of gaps that exist and their length. With that we were able to plot a histogram of the gap size per each one of the series: internal temperature, external temperature and electricity use. This is shown in Figure 1.

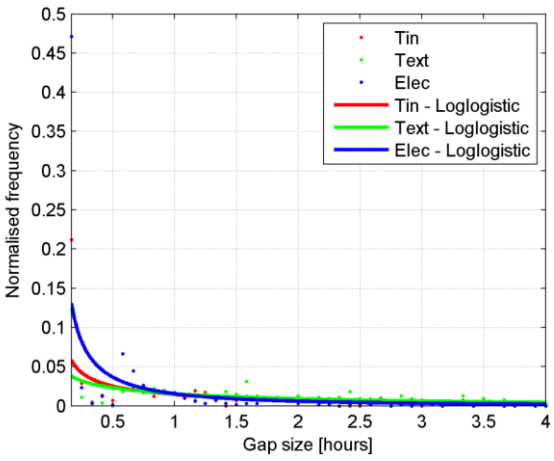

Figure 1 - Gap length normalised frequency and Log-logistic probability density function. The figure starts at x=1/6 what corresponds to a 10 minutes gap length, the minimum gap length in a time series with a 5 minutes sampling period.

When analysing the data from the CHP study collected by the carbon trust, we have seen that most of the gaps

are single points, but one can also see gaps of more substantial length. Thirty buildings were used to study these gaps in each data set. The gaps extracted from the data of each building were added together to have a larger sample of gap lengths that are likely to occur in each case.

Analysis of the data from the ENLITEN project

The ENLITEN project has the aim of quantifying the potential of energy reduction in dwellings by the use of an intelligent energy advisor. As this energy advisor is intended to be placed in building without having a large impact, the project has always had the motivation of using the minimum number of sensors possible. That brought the need of developing a measuring equipment in-house. It was therefore expected failures of the sensors and therefore frequent gaps in the data. The number of houses used from the ENLITEN data to obtain the number and length of gaps was fifty six.

The normalised histogram of the gaps found on the ENLITEN data can be found in Figure 2.

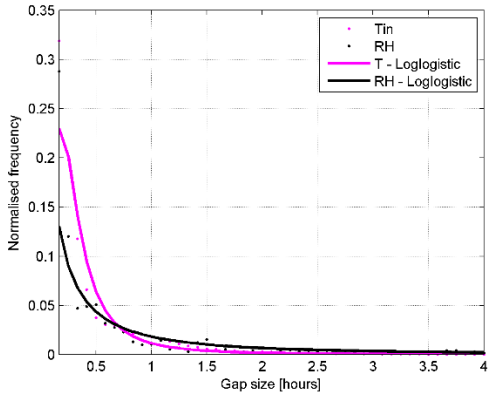

Figure 2 - Gap length normalised frequency and Log-logistic probability density function. The figure starts at x=1/6 (intersection between x and y axis) corresponds to a 10 minutes gap length, the minimum gap length in a time series with a 5 minutes sampling period.

Maximum likelihood estimation of the distribution of the length of gaps

The study of the gaps that are present in real data coming from buildings allowed estimating the probability distributions that indicates the likelihood of having certain gaps in data, their frequency and length.

Several probability distributions were studied analytically and were reduced to three as the most likely to represent the probability distributions of the gaps. These three were the gamma distribution, the exponential distribution and the log-logistic distribution.

Figure 3 shows the comparison of three pdfs that were likely to be able to represent the distribution of the gap length together with the empirical probability distribution of the real data.

To evaluate which one of this distributions was most likely to represent accurately the distribution of gaps, we used the built-in Matlab function *fitdist*.

The routine *fitdist* takes a set of values and a given probability distribution and looks for the parameters of that distribution that would make the likelyhood of the data being generated by that distribution maximun. This is called parameter estimation using maximun likelihood (more about maximum likelihood estimation can be found on (Hamilton 1994)).

This was done with the three distributions mentioned above. The cummulative distributions of each one with the optimised parameters can be compared in Figure 3 with the empirical cumulative distribution from the real data.

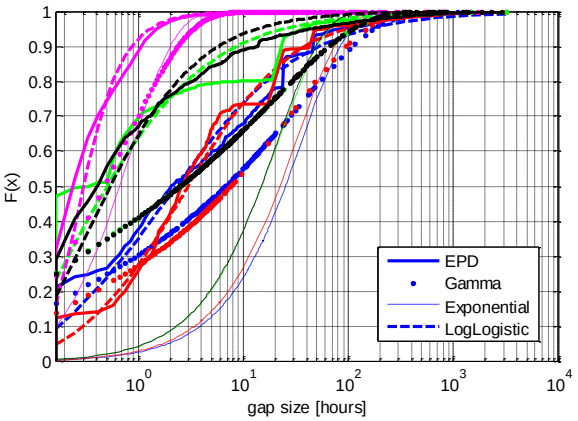

Figure 3 - Comparison of three different pdf's with the empirical cumulative distribution of the gaps of the five variables. There are three sets of curves with three different colours: Blue-TinCHP, Red-ToutCHP, Green-ElecCHP, Mag.-TinENL and Black-RHENL. The line type shows the distribution as in the legend for TinCHP. The curve corresponding with the log logistic of the magenta line is the same as that one of the black line.

Figure 3 shows the empirical cummulative distributions, and the gamma, exponential and log-logistic that fit the data best using the MATLAB built in function fitdist this function uses maximun likelihood optimisation to find the paramters of the distribution that better fits the points except for the lognormal distribution, the estimated value of the sigma parameter inthis case is the square root of the estimate of the variance of the log of the data. The five different colours represent the five different variables (see caption). One can see in this figure that the log-logistic function seems to be the most adecuate to represent the probability of the gaps coming from this environmetal data.

Table 1 shows the parameters of the probability distributions. This table shows how the estimators have been obteined with a large certainty, as the

elements on the covariance matrices of the estimators are order of magnitude smaller than the estimators.

Figure 4 shows a scatter plot of the frequency of the gaps (notice that the length of the gaps is a discrete variable as the series is sampled) and the representation of the three probability distributions on a log-log plot. This graph has been included in this paper to outline the fact that the probabilities of the occurrence of the gap flattens for large gaps.

Table 1 - Parameters of optimal Log logistic distributions for the CHP data.

	log location	log scale	Cov. matrix of estimates
Tin	3.29	1.44	1.5e-2 -3.7e-5 -3.7e-5 3.2e-3
Text	3.54	1.18	7.6e-3 4.1e-5 4.1e-5 1.7e-3
Elec	1.63	1.36	8.2e-3 5.6e-4 5.6e-4 1.8e-3

Table 2 - Parameters of optimal Log Logistic distributions for the ENLITEN data.

	log location	log scale	Cov. matrix of estimates
Tin	0.968	0.585	1.8e-5 6.4e-7 6.4e-7 3.9e-6
RH	1.24	0.621	2.9e-6 7.3e-9 7.3e-9 5.7e-7

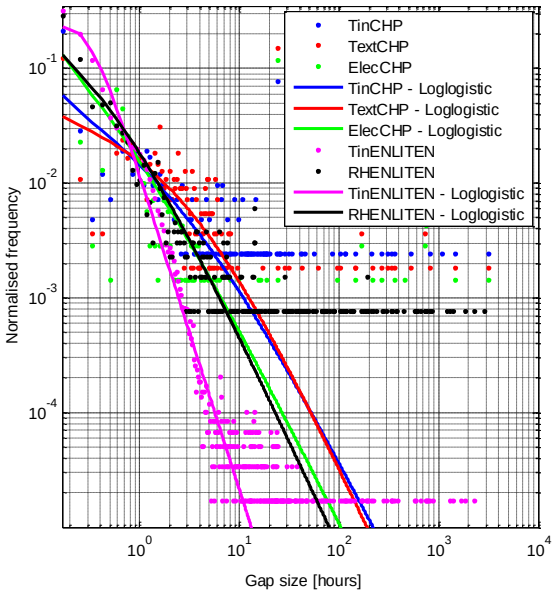

Figure 4 – Scatter plot of the occurrence of the gaps with respect to their length.

Reconstruction methods

Six methods were used for the interpolation of the data. The first four are simple methods and use only the data points boundding the gap to calculate the points missing.

The first one is the '**nearest**' or zero order algorithm. This algorithm assigns the value to the missing point of the closest point that exist in the series. It is a very conservative method as it never takes values that the series does not show in other points.

The second algorith uses **lineal** interpolation. With this method, a straight line is computed between the data points that bound the gap. The missing data points are assigned the corresponding values according to this line.

Spline was the third method used. In this case the interpolation uses the derivatives of the datapoints at the bound of the gap to ensure that the interpolation produces a smooth result.

The **cubic** reconstruction is the fourth method. When this interpolation is used the interpolated valuse are based on a shape-preserving cubic interpolation of the values that bound the gap.

For the fifth method an **ARIMA** model was created of the time series and then used to forecast the data in time series format. The ARIMA models attemp to capture the patterns of time series by studying the reslationships between each datapoint and those preceding it. This method of interpolation is substantially more complex that the ones shown above, but it was expected that it will outperform those for large gaps.

The ARIMA models are used when correlation is seen between present values in the time series and precedental values. In our case we have used the ARIMA models to model the time series, once the model is fit to each data set, the missing values are forecasted. For forecasting, the noise of the ARIMA model is made zero so the value given for the forecasting is the mean of posible values in each step.

To find the ARIMA models of each time series the MATLAB build-in function *estimate* was used. This fucntion uses maximun likelihood to find the optimal parameters of the model to represent the given data.

It was seen that the ARIMA topology for temperatures and electricity were fundamentally diferent. This is unsurprising as the series have very different behaviour.

After a series of trials we found that the right ARIMA model for temperature and humidity was of the form in Equation 1.

$$(2, 0, 1) \times (3, 0, 0)_{24} \quad (1)$$

This model can be represented in terms of the componetns of the series. This is shown in Equation (2):

$$y_t = \mu + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \theta_{t-3} y_{t-3} + \Phi_1 y_{t-24} + \Phi_2 y_{t-25} + \Phi_3 y_{t-26} + \varepsilon, \quad (2)$$

where 'y_t' is the value of the variable at time 't', 'φ_k' is the coeffienct for autoregressive term 'k', 'θ_k' is the coeffienct for moving average term 'k', 'Φ_k' is the coeffienct for the seasonal autoregressive term 'k' and 'ε' is the term representing the random normal noise of the process.

For electricity, the ARIMA model topology that was found to represent the series best can be seein in Equations 3 and 4.

$$(1, 0, 0) \times (3, 0, 0)_{24} \quad (3)$$

$$y_t = \mu + \varphi_1 y_{t-1} + \Phi_1 y_{t-24} + \Phi_2 y_{t-25} + \Phi_3 y_{t-26} + \varepsilon. \quad (4)$$

with nomenclature in agreement with Equation 2 and 3.

The sixth method used is based on reconstruction of the series in the frequency domain. In this case, the **Extended Discrete Fourier Transform** (EDFT) was used to obtain the spectrum of the signal. When the time series have gaps (as it is the case in reconstruction), the EDFT uses an optimisation loop that adjusts the value of the Fourier Transform in each frequency in a way that minimises the difference between the time series created converting the fourier transform to the time domain and the original on the points where this has values.

The implementation used for calcualting the EDFT was (Liepins).

As the fifth and sixth methods are more computationally expensive and likely to be the ones used for larger gaps we down-sampled the data from 5 minutes to 1 hour sampling period.

This was done in the following way: first the series was smoothed with a sindow of one hour and then the smoothed series was resampled. With this, we avoid trying to represent high frequency noise with the reconstruction method.

APPLICATION AND RESULTS

Selection of the series and the gaps

After observing the gaps that may occur in environmental data five gap lengts were selected to make sure that all pausible scenarios were investigated. This is summarised in Table 3.

Table 3
Gap lengths to be tested

Gap size [hh:mm]	Equivalent number of Missing data points
00:10	1
00:30	6
01:00	12
03:00	36
06:00	72
12:00	144
24:00	288
72:00	864

The selection of the gaps lengths has been such to represent the whole range of gap lengths that one can see in reallity according to the first section of this paper.

To evaluate the reconstruction methods five real and complete time series were selected with no gaps. These series had a length of 30 days due to the significance of this period in environmental data (~one month). To evaluate the reconstruciton gaps were put on them artifitially. Gaps will be created in these series on a random location, and only one gap at a time. The series are shown in Figure 5.

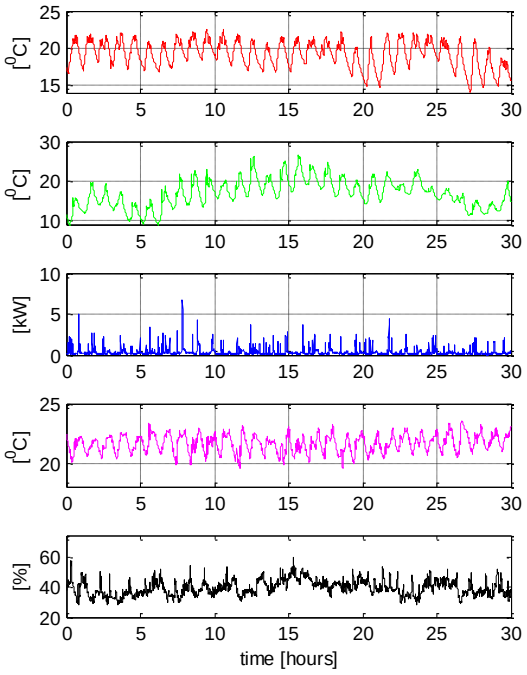

Figure 5 - Series used for assessment of the reconstruction methods. Internal temperature CHP (red), External temperature CHP (green), Electricity (blue), Internal temperature ENLITEN (magenta), Relative humidity ENLITEN (black).

Interpolation of the internal temperature from the CHP project

As seen in Figure 5, the temperature from the CHP data used for calibration is rather smooth, and has a substantial periodicity with little noise.

The interpolation of ten random gaps of the sizes given in Table 3 has been done for each of the 5 methods described before. To calculate the accuracy of the interpolation the summation of the squares roots of the sum of the squares of the residuals were used. Although this is a commonly used estimator of the quality of the fit, it has been seen that visual inspection is in general more powerful than this to evaluate if the method was capable of capturing the dynamics of the problem.

Figure 6 shows the mean of the residuals of the ten interpolations performed each one in a diferent location, and also the ranges showing the 95% intervals calculated using the standard deviation of the ten runs.

One of the first things that can be seen in this graph is how all methods except the EDFT are more accurate as the gap size decreases. This can be explained due to the fact that the EDFT reconstructs the series effectibly in the frequency domain.

It can also be seen that splines should never be used to interpolate this kind of data as it can be a rather unstable method and produce very inaccurate interpolations as the gap size become larger.

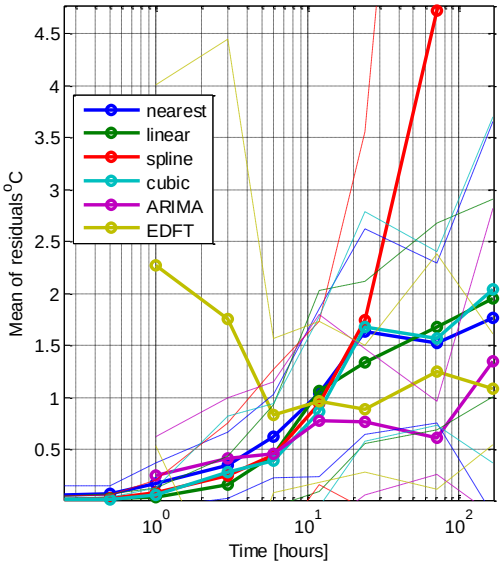

Figure 6 - Residuals for the interpolation of internal temperature from CHP data. Solid lines and markers are the mean of the residuals from the ten runs. Dashed lines are the 95% intervals considering the deviations between the ten runs.

This graph also shows how any gap that is interpolated and has a size larger than twelve hours should be interpolated using one of the more complex methods either ARIMA or EDFT, for gaps smaller than this length one could use indistinctly any of the other methods but never the EDFT.

It can be seen that gaps of 6 hours show the same error with lineal, cubic spline and ARIMA interpolations. It seems adequate to interpolate those gaps with the simpler models namely lienal, spline and cubic, as ARIMA would take much longer and the result would be similar.

The computation times have been shown in Figure 7. This graph shows how the computational time of the basic interpolation methods is rather trivial and therefore these methods should be used for small large numbers of gaps.

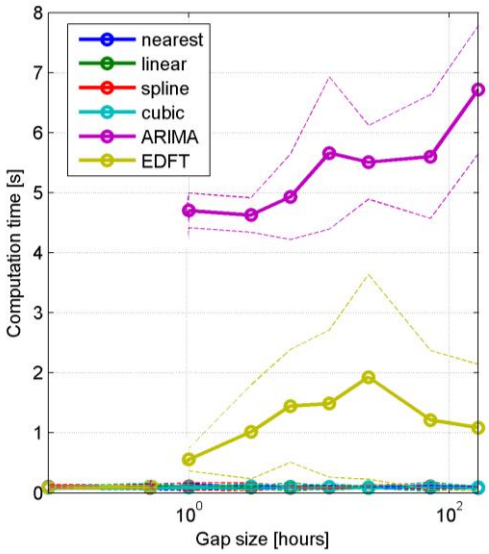

Figure 7 - Computational times of the different methods for different computational times.

The ARIMA method is the most computationally expensive, as it needs to find the right ARIMA model of the data to do the interpolation. The EDFT method takes less time than the ARIMA method, and also, this time is reduced with the gap size whereas with the ARIMA method the time increases slightly. It should be noted that for these two methods the computational time is not multiplied by the number of gaps that need to be filled, instead, one calculation will be sufficient to have the information to fill all gaps in the series.

Interpolation of the external temperature from the CHP project

The external temperature series is similar to the internal temperature of this data set. This is again a series with large periodicity (seasonality in time series jargon) and it is rather smooth.

The residuals of the interpolations can be seen in Figure 8.

This graph shows again the strength of the interpolation using the ARIMA method. This is the best method in any gap larger than 6 hours.

As before, gaps that are 6 hours long or shorter should be recnostructed with other methods. In this case the cubic interpolation seems marginally better than the lineal and spline. This method may be then be the most adequate ehn interpolating temepratures for gaps smaller than this threshole whereas larger gaps should be interpolated with ARIMA.

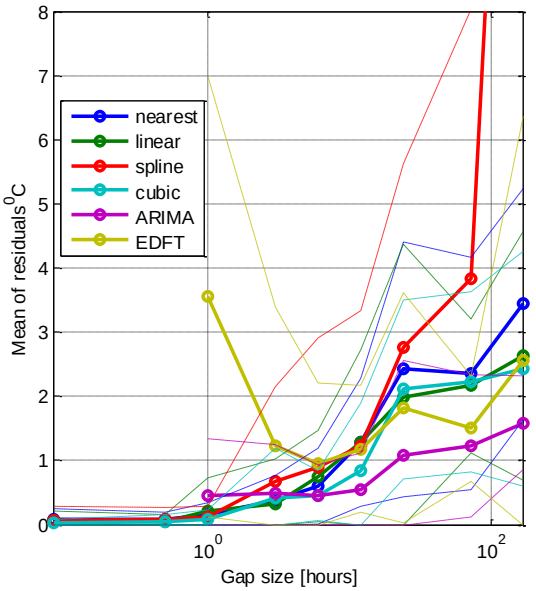

Figure 8 - Residuals of the interpolation of the external temperature using CHP data.

It has been seen in this series that the EDFT method may not be safe as it might get some of the harmonics incorrectly calculated if the amount of data available is reduced. This was not seen with the ARIMA method. Figure 8 shows the average of the value of the residuals for each method, but also the 95% confidence intervals created with the standard deviations of the interpolation of ten gaps located in aleatory positions.

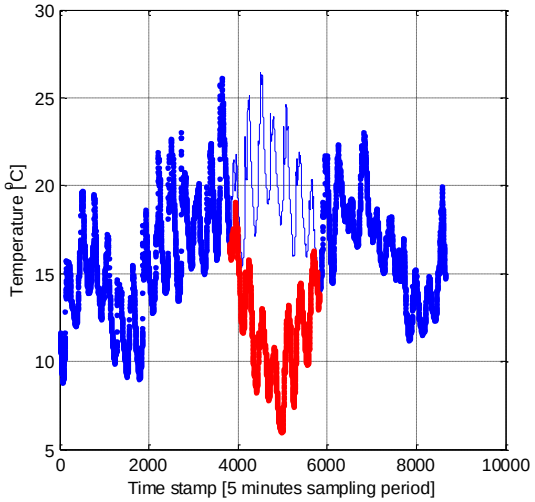

Figure 9 - Interpolation done using the EDFT method. The dotted line is the data used as known data for the interpolation. Interpolation (red).

It can be seen that this interval gets rather large for the EDFT method, and that is because of the effect previously mentioned. An example of this is shown in Figure 9.

This figure shows how the EDFT calculated that there must be an harmonic that create a rather larger oscilation in the gap introducing a large error. This can be very dangerous and visual inspection should be used when applying this method.

The graph with the computational times has not been included for this variable as it is very similar to the one shown in Figure 6.

The computational times show in Figure 6 can be used for the rest of the study. Although the ARIMA method and the EDFT are intrinsicly optimisaiton methods that may take different times depending on the data, we have seen that the time differs very little from variable to variable.

Interpolation of the Electricity use from the CHP data

The electricity data is in general noisier than temperatures and humidity. This is because it is not the response of a physical system with a given inertia, but just the response of the occupant behaviour and operation of machines. It was expected when selecting this series that the interpolation would be more challengin and foundamentally different.

So much is so, that the topology of the ARIMA model had to be changed to ensure that the time series were represented properly, as decribed in the previos section.

The residuals of the interpolation errors can be seen in Figure 10.

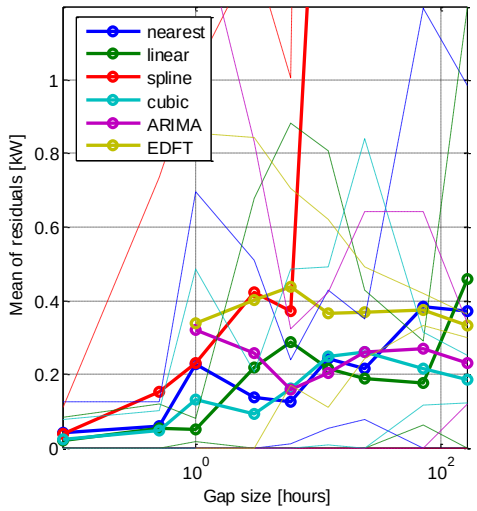

Figure 10 – Residuals of interpolation of Electricity from CHP data.

The figure shows what we anticipated. The interpolation of electricity use is a different problem. Although some periodicity is found, the large spikes are difficult to model leading to larger errors in interpolation. This does not come as surprise as previous literature in the topic has shown that more accurate ways of replicating electricity use are Markov chains (Richardson and Thomson 2010), method that has little to do with the methods we have used here.

The reader is recommended to use those methods when interpolating this kind of series.

If one of the methods proposed here had to be selected for this interpolation, that would be again the cubic, as this method seems to have smaller erros in average but also smaller variability of residuals what means that is less risky to use this interpolation. To have a good interpolation, only very small gaps should be interpolated. Gaps representing one or two missing datapoints.

Interpolation of internal temperature from ENLITEN dataset

The internal temperature from the ENLITEN dataset was also used to evaluate the interpolation methods. The results can be seen on Figure 11.

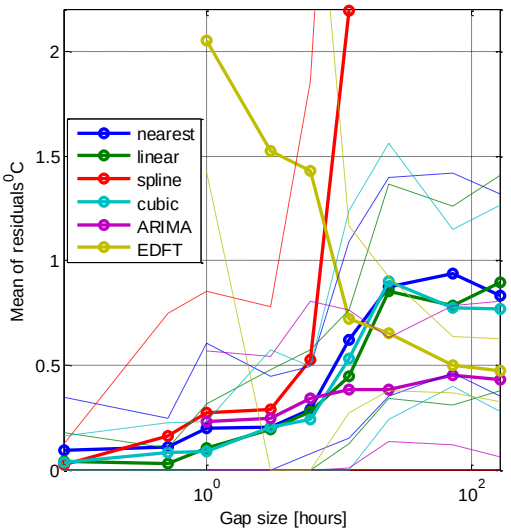

Figure 11 – Residuals of interpolation of internal temperature with data form ENLITEN.

Figure 11 shows consistancy with the results found when interpolating the internal temperature with data from the CHP study what suggests that the findings may be independent on the data set. The simpler methods interpolate well the small gaps (<6 hours) whereas EDFT work well for larger gaps. It is even more clear in this interpolation that the EDFT may not be adequate for small gaps. Whereas ARIMA seems to be preferable for long gaps as was also seem in Figure 6.

The 6 hours threshold is also seen in this dataset. After the 6 hours gap the ARIMA method is substantially better than the simple methods, it is only for those gaps that this method should be used at the cost of longer computational times.

Interpolation of relative humidity from ENLITEN data

The relative humidity is rarely available in indoor environmental data in buildings. However it is a very important variable and can be highly linked with air quality and therefore comfort. The series of humidity are more "spiky" what can make the interpolation difficult. However, it does present a large periodicity what may make one think that a certain level of interpolation is possible.

The residuals of the interpolations of the relative humidity data from the ENLITEN project are shown in Figure 12.

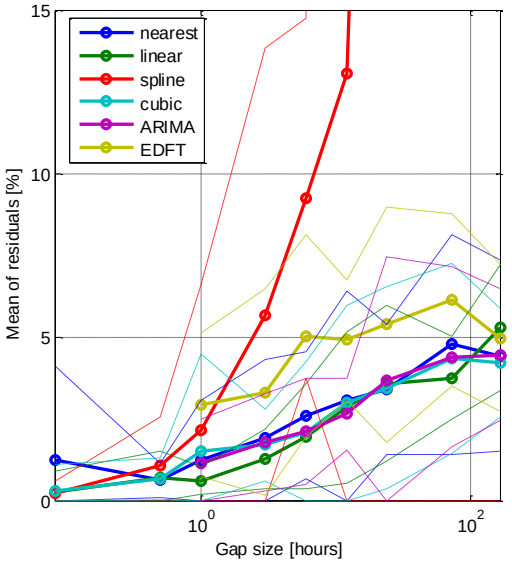

Figure 12 - Residuals of the interpolation of relative humidity from ENLITEN data.

The residuals of the interpolation for relative humidity are substantially different to those of the temeprature either if it is internal or external.

The large separation between the simple and the elaborate methods found after the 6 hours gap can not be seen when interpolating relative humidity. Instead, the results show that the error grows with gap size despite the method, and that any of the simple methods except the spline will perform as well as the ARIMA method. The EDFT method seems to have the precision of the other methods only for gaps of three days. It is plausible that the EDFT outperform the other methods when gaps are larger than three days when one ovserve the reuslts but this would nbeed to be evaluated in future work.

With the results obtained so far, the ideal method would be the lineal or cubic method as it has the lower computational time without compromising accuracy.

CONCLUSION

This work shows a study on gaps existing on data coming from measured data in buildings. We have shown that gaps in data from buildings seem to follow a log-logistic distribution; this finding can be valuable for researchers to anticipate to the interpolations problems that they may have when using real data.

We have also seen that these gaps follow the same distribution for very different variables, as they are internal temperatures from a sensor in the house and electricity from an amperimetric clamp at the counter

of the building. In addition, the same distributions have been found good for data sets coming from different projects with different hardware.

Six interpolation methods to reconstruct missing data in time series have been evaluated. The methods compared can be separated in two groups: the local methods that only use the data points that bound the gap, and the second group that looks into the full series to infer the missing data.

We have seen that interpolation always come with an error, but on this work, we show that one may want to use a different method depending on the gap length and the time constrains. The limit of 6 hours seems to be key when one want to interpolate time series coming from building data.

The results suggest that gaps smaller than 6 hours can be interpolated with simple methods based in interpolation without losing accuracy whereas gaps larger than that should be filled with forecasting of previously estimated ARIMA models.

This statements hold truth with the series of temperature (external and internal). With relative humidity, it has been seen that local simple methods are methods that are as good as more complex and therefore is not worth to use the more complex ones as they will result on longer computational times without improvements in accuracy.

For the case of electricity, none of the methods tested in this work would be adequate. Instead, the operator should either transform the series. For example, calculate the logarithm of the series as done in time-series analysis in some cases or use other methods such as Markov chains.

With this work, we do not pretend to suggest or find the perfect method to fill the gaps in environmental data from buildings, mainly because we believe there is not such a thing. However, we believe that the findings from this work could be useful when practitioners and researchers fill obliged to interpolate data to feed it in controls, simulators and so forth.

ACKNOLEDGMENTS

This research has been performed in the project ENLITEN (Energy literacy through an intelligent home energy advisor) funded by the EPSRC (EP/K002724/1).

REFERENCES

Coley, D. A. and J. M. Penman (1992). "2nd-Order System-Identification in the Thermal Response of Real Buildings .2. Recursive Formulation for Online Building Energy Management and Control." Building and Environment **27**(3): 269-277.

Crawley, D. B., J. W. Hand, M. Kurnmert and B. T. Griffith (2008). "Contrasting the capabilities of building energy performance simulation programs." Building and Environment **43**(4): 661-673.

Eames, M., T. Kershaw and D. Coley (2011). "On the creation of future probabilistic design weather years from UKCP09." Building Services Engineering Research & Technology **32**(2): 127-142.

Hamilton, J. D. (1994). Time series analysis, Princeton university press Princeton.

Liepins, V. "Extended Discrete Fourier Transform." from http://www.mathworks.com/matlabcentral/fileexchange/11020-extended-dft.

Madsen, H. and J. Holst (1995). "Estimation of continuous-time models for the heat dynamics of a building." Energy and Buildings **22**(1): 67-79.

Richardson, I. and M. Thomson (2010). Domestic electricity demand model - simulation example, Loughborough University