A Functional Principal Components Model for Internal Loads in Building Energy Simulation

Rebecca Ward[1], Ruchi Choudhary[1], Yeonsook Heo[2], John Aston[3]
[1]Energy Efficient Cities initiative, Department of Engineering, University of Cambridge, UK
[2]Department of Architecture, University of Cambridge, UK
[3]Statistical Laboratory, University of Cambridge, UK

Abstract

There is currently no established methodology for quantifying uncertainty in occupant-related building internal loads. In this paper, we propose that distinct spaces within a building may be assigned an operational signature comprising the daily base load, load range and diversity profile. A Functional Data Analysis (FDA) approach has been used to analyse monitored electricity consumption data for the derivation of such signatures. This approach enables simulation of the inherent stochasticity. It represents a step forward towards an ability to propagate uncertainty through a building energy simulation and to quantify the change in electricity consumption associated with a change in building operation.

Introduction

Schedules of internal loads in a building are difficult to quantify in a way which accurately represents existing demand, yet also permits estimation of the impact of changes in building operation. In a typical building energy model loads are characterised by user-defined energy use intensity and diversity profiles per end-use (e.g. lighting, plug loads, large appliances); this approach is straightforward to implement and interpret. However, it can lead to large differences between simulated and actual energy use (de Wilde, 2014) and as yet there is no established methodology for reconciling those differences. Moreover, it is difficult to quantify the uncertainty associated with the input and propagate that uncertainty through the simulation. Internal loads are inextricably linked to building occupancy, and a considerable body of research is currently being performed in order to understand occupancy patterns using data-mining and machine-learning techniques (Liang et al., 2016; D'Oca and Hong, 2015; Khosrowpour et al., 2016), to explore different ways of modeling occupancy schedules (Feng et al., 2015; Wang et al., 2016) and to implement models of the energy consequences of different occupant behaviours (Mahdavi and Tahmasebi, 2015; Yan et al., 2015). The study by Mahdavi et al. (2016) highlighted the need for better models of plug loads for building performance simulation and explores detailed monitored occupancy and plug load data for a selected office. However while data derived from in-depth empirical monitoring studies are useful, in the commercial environment such occupancy data are not typically available; as a consequence it is useful to consider whether an effect-based approach using available monitored electricity consumption data might be viable.

A recent study explored the available alternative methods for modelling internal loads concentrating specifically on small power demand or 'plug loads' (Ward et al., 2016b). The study highlighted the need for a bottom-up stochastic approach to be used if the key parameters of interest are related to the timing of the daily variation in electricity consumption. Yet bottom-up models may be time-consuming and difficult to verify due to the large numbers and variety of devices used in buildings (Rysanek and Choudhary, 2015; Menezes et al., 2014). It seems pragmatic to use available data to inform load models wherever possible, particularly monitored electricity consumption data, and top-down data-driven approaches have also been investigated (Sun, 2014; Wang et al., 2016). These approaches are useful for calculation of aggregated internal load profiles, but cannot be used to attribute energy demand to particular devices or areas of buildings, or to identify the sources and timing of large and/or inefficient energy consumers.

In this paper we propose an alternative method. Building on the approach used in the UK National Calculation Methodology (Building Research Establishment, 2016), we propose that distinct spaces within a building may be characterised by the activities of the occupants and therefore may be assigned a distinct *'operational signature'* comprising the base load, daily range of power demand and a diversity function. A Functional Data Analysis (FDA) approach has been used to analyse monitored electricity consumption data for the derivation of such signatures for an existing building. The novelty of the proposed model lies in the definition of an operational signature per space in a way which separates the base load from the load range. It facilitates simulation of the inherent stochasticity and hence the uncertainties in the occupant-related loads. Uncertainty in base load and load range may be quantified directly from the monitored data. Quantification of uncertainty in diversity, however, is challenging owing to the representation of diversity as a function as opposed to a series of independent hourly values. Functional Principal Component Analysis (fPCA) is a statistical approach that enables a description of the variation across a time period and permits a mathematical interpretation of the shape of the diversity functions for different use types. It can also lead to a stochastic representation of the diversity which may be used to quantify the associated uncertainty.

This paper builds on initial studies presented previously (Ward et al., 2016a), but here a more efficient mathematical technique has been used; the diversity functions are separated into phase and amplitude elements in order to perform fPCA on each element (Marron et al., 2015). fPCA of the phase is termed *'Horizontal fPCA'* as it represents the variation in time, or x-axis variability, while fPCA of the amplitude is termed *'Vertical fPCA'* as it represents the variation in magnitude, or y-axis variability. Random sampling of principal component coefficients for both elements prior to recombination yields a stochastic representation of the diversity function. This is an efficient and powerful new technique, particularly applicable when the phase of the data is important, as in this case.

The following section of the paper outlines the methodology and illustrates why separation of the data into phase and amplitude is relevant in this case. The building and the electricity consumption data used for the analysis are also described and a necessarily brief outline of the mathematical framework for the analysis is provided. Subsequent sections show the application of the methodology to the data of interest and the results in terms of sample diversity curves. The model has been assessed against its ability to predict two Key Performance Indicators (KPIs), namely the timing of the daily peak load demand and the daily total electricity consumption. The discussion and conclusions provide a critique of the approach and assess the benefits of this model over models currently in use.

Methodology

A fundamental aspect of the model developed here is that it is based on data and thereby lends itself to further refinement and expansion as more data become available. To this end, electricity consumption data from the William Gates building in Cambridge, UK, have been used as a basis for the model. This building, completed in 2001, houses the Cambridge University Computing Laboratory. It is constructed on 3 similar floors, and is sub-metered at hourly intervals for plug loads and lighting at a fairly high spatial resolution. Spatial zones are identified according to the building orientation (Rice et al., 2010; Ward et al., 2016a). The spatial zones are also typically occupied by different types of user as indicated in Table 1.

Table 1: William Gates building, zone use types

Zone	Ground	First	Second
East	Admin	Student	Student
South-East	Mixed-use	Mixed-use	Mixed-use
South-West	Canteen	Classroom	Classroom
West	Classroom (Lecture Theatre)	-	-
North-West	Admin	Meeting Space	Classroom
North-East	Admin	Student	Student

An in-depth analysis of the data indicated that zones with similar use types exhibit similar diversity profiles that are different from those for a different use type (Ward et al., 2016a). Consequently, four typical zones were selected for this analysis representing one of each type of use, namely:

- Student - East, 2nd floor (E2)
- Administration - East, ground floor (EG)
- Mixed-use - South-east, 2nd floor (SE2)
- Classroom - South-west, 2nd floor (SW2)

For the purposes of input into a building energy simulation it is common practice to parameterise the power demand by the daily peak value and an hourly diversity profile in which diversity is defined as the fraction of the peak power demand (Wilkins and Hosni, 2000). However, we suggest that the base load is not best modelled as a fixed proportion of the peak load and hence we propose an alternative parameterisation by base load, load range (i.e. peak load - base load) and diversity. In this case the diversity is a factor on the load range and varies from 0 to 1. Figure 1 illustrates the plug loads for term-time weekdays in the academic year 2013-2014 for two student zones, E2 and NE1. The top figure shows the normalised data, with the base load, load range and mean diversity factors below. As can be seen the two zones are quite different, particularly in terms of the base load, yet are similar in terms of the mean diversity. By comparison, the National Calculation Methodology suggests a diversity factor ranging from 0.05 to 1 on a peak load demand of $11.99W/m^2$ for a cellular University office (Building Research Establishment, 2016). This yields the same base load for both zones of $0.6W/m^2$ which is not only too low but also does not adequately capture the difference between the zones.

Separation of phase and amplitude

Considering the diversity we suggest that the differences between use types can be described in terms of the start and end times of the day, the rate of increase/decrease at the beginning and the end of the day and the timing of the daily peak. So the time - or phase - variation of the data is important. Marron et al. (2015) state that *'the presence of phase variation can play havoc with classical data analyses that are designed for data structure without phase changes'*, and discuss the requirement of registration

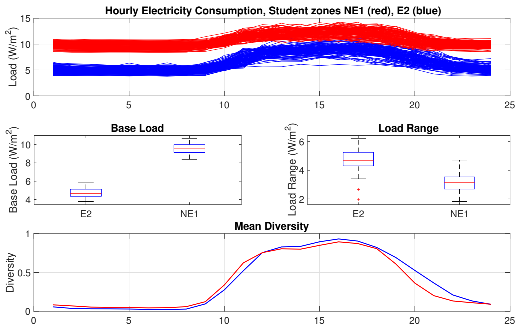

Figure 1: Comparison of student zones

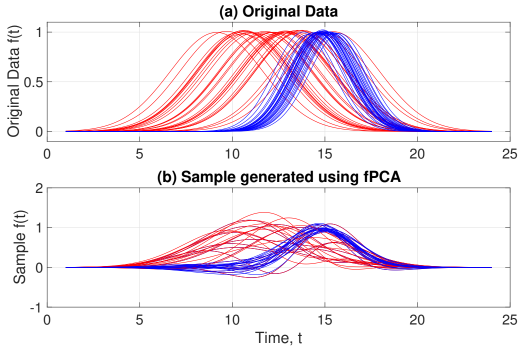

Figure 2: Example data and standard fPCA samples

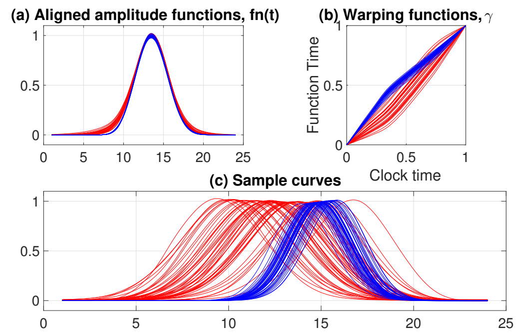

Figure 3: Samples generated using phase and amplitude separation

of curves prior to such analysis, effectively removing the phase relationship by aligning the curves. In our previous study (Ward et al., 2016a), a functional Principal Component Analysis (fPCA) of the diversity data was performed deliberately without prior registration of the data as it is the daily phase relationship that is of primary interest. Following the fPCA approach, the diversity profile, $\boldsymbol{D}$ for each activity type, i was expressed as a linear combination of the mean diversity profile, $\boldsymbol{\mu}$ across all activities and a weighted sum of the dominant modes of variation from the mean in the following manner:

$$\boldsymbol{D_i} = \boldsymbol{\mu} + \sum_{j=1}^{N} \alpha_{i,j} \boldsymbol{\nu_j} \tag{1}$$

where the N principal components, $\boldsymbol{\nu_j}$ are summed according to the weightings or 'scores', $\alpha_{i,j}$. In this equation, the diversity, mean and principal components are all functions. Sample diversity functions were then generated by random sampling from the distributions of scores for each type of activity.

The problem with this approach is that if the curves are not aligned initially the mean profile and principal components that are calculated are averaged over time and the difference between distinct peaks becomes blurred. In this case generation of random samples by sampling principal component scores can lead to unrealistic diversity curves.

As an example of this, consider the synthetic dataset illustrated in Figure 2(a). This dataset contains two distinct groups of data, mimicking our different zones; it has been generated using the equation $f_i(t) = z_i e^{-(t-a_i)^2/2}$ where the z_i are randomly sampled from a normal distribution $N(1, 0.01)$ giving only a small variation in peak amplitude. The variation in phase i.e. the timing along the x-axis, is much greater; the parameter governing the phase relationship, a_i, is generated in two groups, one from $N(12, 1.5)$, plotted in red, and the other from $N(15, 0.5)$, plotted in blue. Sample curves generated from a functional Principal Component Analysis without prior registration of the data are shown in Figure 2(b). Mixed modes are visible, i.e. curves that don't exhibit the same general shape as the original data, and while the samples do show similarities to the original data - particularly the mean location of the peak - there is a wider variation in amplitude than in the original data. By comparison consider Figure 3. The approach to be used in the rest of this analysis has been applied to this dataset; the synthetic functions can be represented by a set of amplitude functions (Figure 3(a)) which may be warped to the original data using the warping functions presented in Figure 3(b). The warping function relates *function time* on the y-axis to the *clock time* on the x-axis; points on the curve above the diagonal imply that the function lags behind the mean aligned function and vice versa. Separate statistical analyses of the amplitude and warping functions followed by a recombination of the two elements has been used to generate the sample curves shown in Figure 3(c). These are clearly more representative of the original data than those shown in Figure 2(b) indicating that this new approach is more applicable to data where the phase relationship is significant.

So how is the separation of the functional data into phase and amplitude performed? In itself, this is an extensive topic of research, but good results have been achieved using an *'elastic shape analysis'* approach, as detailed in Srivastava and Klassen (2016). Elastic shape analysis of the functional data *f(t)* relies on the calculation of the Square Root Slope Function, or

SRSF, *q(t)*, defined by:

$$q(t) = sign(\dot{f}(t))\sqrt{(|\dot{f}(t)|)} \tag{2}$$

It is desired to extract warping functions, $\gamma(t)$ that map the original data functions $f(t)$ to a set of aligned functions $fn(t)$. However, if we consider two such data functions, f_1 and f_2 and a warping function γ (omitting the domain for clarity), in general it is the case that,

$$\| f_1 - f_2 \| \neq \| (f_1 \circ \gamma) - (f_2 \circ \gamma) \| \tag{3}$$

The problem with this is that the optimal alignment of the two functions will differ according to the direction of alignment. This may not be an issue if the warping functions themselves are not of interest i.e. if registration of the curves is the only aim. However, if the analysis of the warping functions is of interest, as here, it is necessary to find a symmetric approach. Using Equation 2 and calculating the SRSF values, q_1, q_2 of f_1, f_2, it transpires that:

$$\| (q_1, \gamma) - (q_2, \gamma) \| = \| q_1 - q_2 \| \tag{4}$$

where $(q, \gamma) = (q \circ \gamma)\sqrt{\dot{\gamma}}$, is the SRSF of $f \circ \gamma$. This isometry property means that it is possible to align the SRSF values by minimising the distance between the aligned SRSFs and a mathematical mean using a Dynamic Programming algorithm, and then to map the aligned SRSF values back to the function space to retrieve the aligned functions, $fn(t)$.

This analysis delivers a set of aligned amplitude functions, $fn(t)$ and a set of warping functions $\gamma(t)$. Functional principal component analysis of both sets of functions may then be performed in order to generate sample datasets. This results in separate linear equations similar to Equation 1 for the two elements. However, it is important to note that the fPCA may not be performed in the original function space for either of the elements; for the amplitude functions, the fPCA must be performed on the SRSF values. For the warping functions, the set of all warping functions is an infinite-dimensional nonlinear manifold which does not lend itself easily to statistical analysis. In order to proceed, the square root of the derivative of the warping function, $\psi = \sqrt{\dot{\gamma}}$, is calculated; the set of all ψ values is a Hilbert sphere, i.e. a unit sphere in Hilbert space, for which the distance between two ψ values is the arc length on the surface of the sphere. This facilitates calculation of a mean warping function, but principal component analysis still cannot be performed directly on the ψ values; instead the mappings of ψ into the tangent space to the mean ψ value - the so-called *'shooting vectors'* - are used (Tucker et al., 2013).

Following fPCA and sampling of the principal component scores, the sample SRSF values are transformed back into function space to generate sample aligned amplitude functions, $fn(t)$ and the sample shooting vectors are transformed back into warping functions, $\gamma(t)$. The final step is to recombine the sample amplitude and phase functions to generate sample functions, $f(t)$.

The full mathematical details of the approach are complex and are detailed in Srivastava and Klassen (2016). Implementations of the approach for R and MATLAB are also available from Tucker (2015).

Analysis framework

The diversity functions extracted from the monitored electricity consumption data have been analysed using the elastic shape analysis approach outlined above, using the framework set out in Tucker et al. (2013). The steps are as follows:

- Compute the optimally aligned SRSFs and the corresponding warping functions.
- *'Vertical fPCA'* - perform a fPCA of the SRSFs and extract principal components and scores for the y-axis variability.
- *'Horizontal fPCA'* - perform a fPCA of the ψ values mapped into tangent space - the 'shooting vectors' - and extract principal components and scores for the x-axis variability.
- Select the number of principal components that encompass the significant variabilities observed in the data for both the horizontal and vertical elements.
- Fit multivariate normal probability distributions to the scores for the different zones for both the horizontal and vertical principal components. Note that it is necessary to retain the correlation between the scores in the model as although there will be zero correlation across the scores for all zones, within a zone there may be significant correlation.
- Generate sample scores by random sampling from the multivariate normal distributions.
- Use the sample scores and principal components to calculate sample SRSFs and shooting vectors.
- Map the sample shooting vectors back to warping functions and SRSFs back to amplitude functions.
- Combine the sample warping and amplitude functions to give sample diversity curves.

These sample diversity curves may then be combined with sample base load and load range vales generated by random sampling from the data, e.g. as presented in Figure 1 for two student zones, to give sample daily electricity consumption curves.

The application of this framework to the electricity consumption data from the selected Gates building zones is presented in the following section.

Data Analysis

The procedure starts with the alignment of the diversity data curves and extraction of the warping functions and aligned amplitude functions, as illustrated

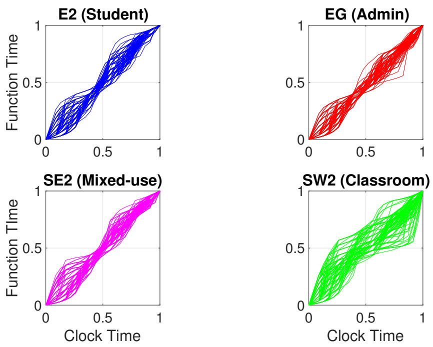

Figure 4: Warping functions

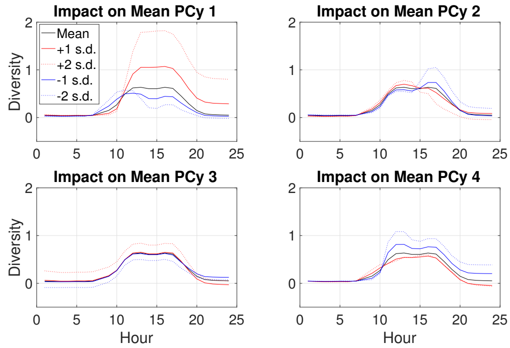

Figure 6: Principal components: amplitude

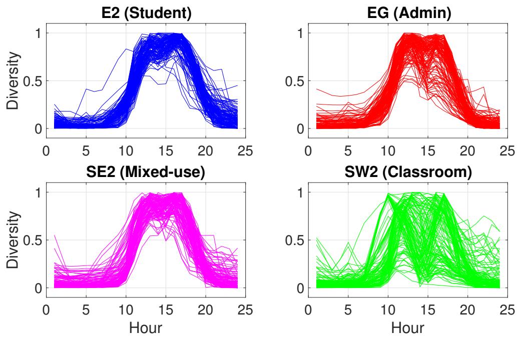

Figure 5: Aligned amplitude functions

in Figures 4 and 5 respectively. At first glance the warping functions look quite similar for the Student and Admin zones, but closer inspection reveals that for the Student zone, the warping functions tend to be higher than the diagonal in the top right quadrant of the plot, which indicates that the diversity curves lag behind the mean later in the day. Similarly for the Classroom, the warping functions tend to lie above the diagonal in the lower left quadrant and below the diagonal in the upper right quadrant, reflecting the tendency for the peaks in diversity to be either before the mean in the morning or after the mean in the afternoon.

There is a considerable amount of variation visible in the amplitude functions after alignment (Figure 5). In conjunction with the variability observed in the warping functions, this indicates that variation in both the amplitude and phase are significant and both must be considered in development of the model.

Having split the data into its constituent parts, the next step is to perform a functional Principal Component analysis for both the amplitude and the phase.

Amplitude

The functional Principal Component analysis for the amplitude is performed on the aligned SRSF values, as described in the Methodology section above. The first 4 principal components are illustrated in Figure 6, where the mean amplitude is shown as the solid black line and the impact of the principal component is shown in colour; a positive score on the principal component is illustrated in red and a negative score in blue. The solid coloured lines show a factor of $+/-$ 1 standard deviation of the score on the principal component, whereas the dotted lines indicate a factor of $+/-$ 2 standard deviations of the score.

The first principal component, PCy 1, is associated with the magnitude of the base load at night and the timing of the morning increase in amplitude, and accounts for 25% of the variability in the data. The second principal component, PCy 2, is associated with the timing of the peak amplitude and accounts for 16% of the variability in the data. The third principal component, PCy 3, is associated with the difference in magnitude between the morning and evening base loads, and the fourth principal component, PCy 4, with the timing of the peak and the difference between peak and evening base load; PCy 3 and 4 account for 13 and 9% of the variability in the data respectively. The higher principal components account for increasingly finer detail of the variability with correspondingly lower significance; increasingly the higher principal components account for noise in the signal.

Phase

The functional Principal Component analysis for the phase yields principal components of the warping functions as illustrated for the first four principal components in Figure 7, but it is more intuitive to understand if we consider the impact of the warping function principal components on the mean amplitude function as illustrated for the same four principal components in Figure 8. The first principal component, PCx 1, is associated with a shift along the x axis i.e. in time and accounts for 30% of the variation in the data, while the second principal component, PCx 2, is associated with a stretch in time and accounts for 14% of the variation in the data. As for the amplitude, the higher principal components account for increasingly finer detail of the variability and ultimately this can be considered as noise.

Analysis of Principal Component scores

Of particular interest here is the extent to which the principal component scores characterise different use

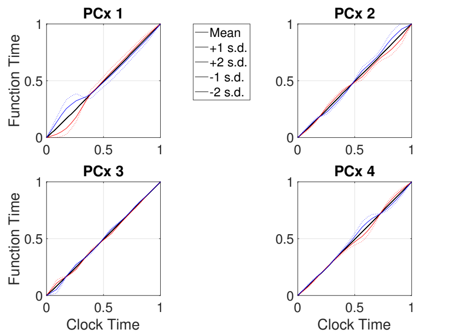

Figure 7: Principal components: phase

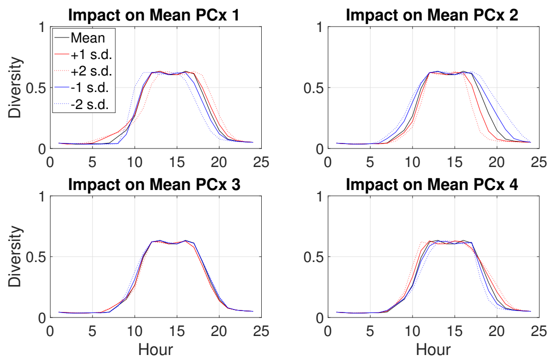

Figure 8: Warping impact of phase PCs

types and separate the different zones. Figure 9 illustrates the principal component scores for the first four principal components of phase (top row) and amplitude (bottom row), where the colour of the dots signifies the zone. For the phase, the plot in the top left quadrant shows PCx 1 on the x axis plotted against PCx 2 on the y axis; there is a clear separation in PCx 2 between E2 (blue), EG (red) and SW2 (green), the student, admin and classroom zones, with the mixed-use zone, SE2 (magenta) lying largely between the student and admin zones. Considering Figure 8 which shows the impact of a positive or negative principal component score on the resulting effect of the warping function, the negative mean score for PCx 2 of the student zone, E2, implies that the diversity is stretched positively in time with respect to the mean, whereas for the admin and classroom zones, EG and SW2, with increasingly positive mean scores, the diversity is stretched negatively - or squeezed - relative to the mean. It is perhaps surprising that less difference between zones is observed for PCx 1, this suggests that the daily variation in start/end times is not as strongly related to the zone as the overall occupancy time of the zone. However, there is a positive correlation between the scores for PCx 1 and 2 for all zones apart from SW2, the classroom zone. This zone does show a difference in the scores for PCx 1 from the other zones, with a negative mean value indicating an earlier finish time as expected for a classroom. There is barely any discernible separation between the zones in the scores for PCx 3 and PCx 4.

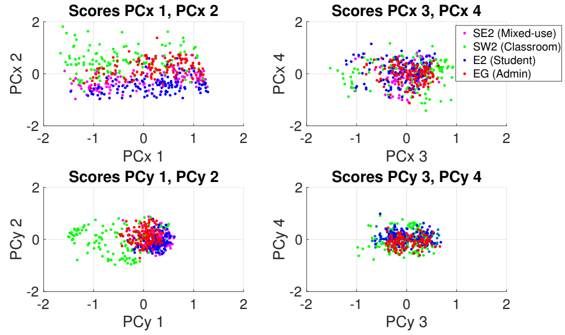

Figure 9: Principal component scores

The bottom two plots in Figure 9 present the scores for the first four amplitude principal components in a similar manner. There is visible separation in the scores for PCy 1 between the different zones and in the scores for PCy 2 between the admin and student zones, EG and E2, shown in red and blue respectively. There is also a small degree of separation in PCy 4 between the admin and student zones. These differences reflect the tendency for the student zone E2 to have an afternoon peak, while the admin zone EG is more likely to have a morning peak. The classroom (SW2) scores for PCy 1 are typically negative, and a comparison with Figure 6 indicates that this reflects the more 'peaky' nature of the classroom diversity.

Results

The principal component scores have been modelled as a multivariate normal distribution for each zone, thereby taking into account the correlation between the zone scores for different principal components. Random sampling from each distribution has then been used to generate score values in order to generate sample warping functions and aligned amplitude functions and these have then been combined to produce sample diversity curves. In line with the aims of this analysis, the number of principal components in the model has been kept to the minimum required to capture the salient variabilities in the data. This has been judged to be two principal components for the phase - capturing the shift and the stretch - and 3 principal components for the amplitude - capturing the timing of the peak and the relationships between the peak and the morning and evening demand.

Sample diversity curves generated using these 5 principal components are illustrated in Figure 11, for comparison against the diversity curves extracted form the original data shown in Figure 10. Visually the variety of the diversity curves illustrated in Figure 11 looks sensible, although there is less variability predicted in the late afternoon for the student and administrative zones (the top two plots in Figure 11) than observed in reality (the top two plots in Figure 10). It is informative to consider the Key Performance Indicators (KPIs) identified as being critical

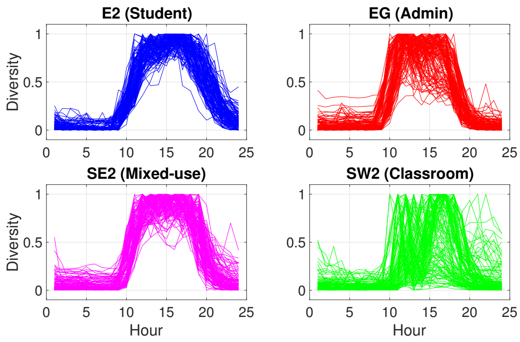

Figure 10: Diversity data

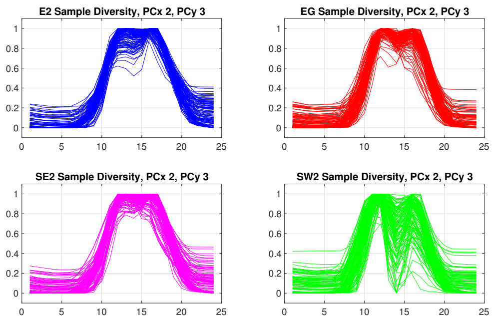

Figure 11: Sample diversity curves

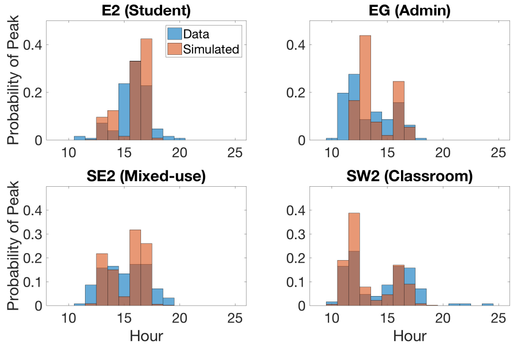

Figure 12: Peak timing

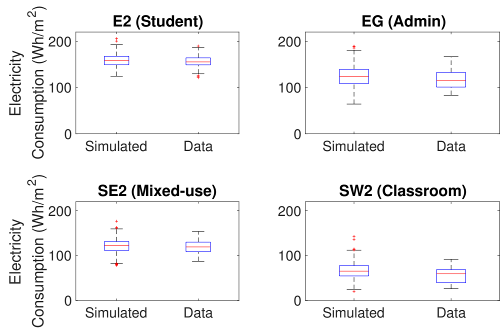

Figure 13: Total daily electricity consumption

to the accuracy of a simulation, namely the timing of the daily peak demand and the daily total electricity consumption. The first of these is purely dependent on the diversity curve, and Figure 12 indicates how closely the model matches the original data. Each histogram indicates for one zone the probability of the peak demand occurring in that hour, with the original data plotted in blue and the sample data plotted in orange. The model results reflect the relative probability of an early or late peak for each zone, but overestimate the peakiness of the distribution. This is due to the small number of principal components used, and further studies will optimise the number of principal components used by exploring which of the principal components are most significant on a zone-by-zone basis.

The second KPI - the daily total electricity consumption - is calculated from three parameters; the base load, the range from base to peak load, and the diversity curve. Figure 13 shows the values for each zone calculated by random sampling from the base and peak load distributions for each zone and the diversity curves calculated using this model. As can be seen, the results of the simulation are in good agreement with the monitored data.

The premise of this model is that the simulated diversity profiles may be used to represent zones of similar use type. To explore this we have considered as a test case the student zone NE1; this zone has a very high base load with a mean of $9.5W/m^2$, and a smaller daily range - around $3.1W/m^2$ - when compared against our sample student zone, E2. However, using the base load and range from the data in conjunction with diversity profiles generated from the model for a student zone yields the results for daily total electricity consumption as shown in Figure 14. These are in good agreement with the data, and the timing of the daily peak is calculated to be at a mean time of 15:52 hrs with a standard deviation of 84 minutes, compared against the monitored data which has a mean daily peak at 15:41 hrs with a standard deviation of 105 minutes - again, a good agreement.

Discussion

Analysis of the data reveals clear differences in the timing of the diversity profile for different users; an ability to capture these differences numerically would engender a step-change in our ability to predict the uncertainty in the timing of the daily peak demand.

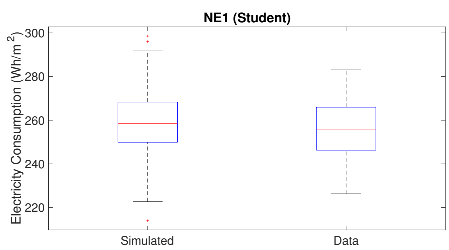

Figure 14: Total daily electricity consumption, Zone NE1 (Student)

To this end, the FDA approach used here has involved separation of the diversity data into amplitude and phase functions followed by fPCA of each element separately. This facilitates a numerical analysis and quantification of the shape of the diversity curve. For example, the principal components of phase primarily relate to either a translation or a stretch in time with respect to the mean, whereas the principal components of amplitude govern the relative amplitude of the morning and evening base loads with respect to the peak load, together with the timing of the peak. Intuitively these components appear sensible, and a consideration of the principal component scores for each use type reveals both the extent to which each component is important for each use type and collectively the extent to which that component might differentiate between use types. For example, the scores for the second principal component of phase, PCx 2, which results in a stretch or squeeze of the diversity curve with respect to the mean, exhibit clear differences for the different use types (Figure 9, top left plot, y axis). This is an important aspect of the difference in diversity between the different use types, and the magnitude of the mean score gives an indication of the reason why; a greater negative mean score for the student zone, E2, plotted in blue, than for the Admin zone, EG, plotted in red, indicates that the student zone is more likely to be occupied for a longer period during the day.

The choice of the number of principal components to include in the model was deliberately made to keep complexity of the model to a minimum. Only 2 principal components for phase and 3 for amplitude were retained, thereby accounting for just 43% of the variability in the phase data and 53% of the variability in the amplitude data. However the results indicate that the model is adequately capturing the behaviour of interest; Figure 12 shows that the difference between the timing of the peak for different use types has been simulated reasonably well, and agreement for the total daily electricity consumption between the simulation and the monitored data (Figure 13) is good. This is true even for the classroom zone which is difficult to simulate due to the highly variable nature of the daily diversity profile. Indeed the sample diversity curves (Figure 11) compare well with the original data (Figure 10). These results are considerably better than would be achieved using a simplistic approach such as the National Calculation Methodology (Building Research Establishment, 2016), with the added benefits of providing an estimate of the timing of the peak for different zone activities and quantification of the associated uncertainty.

This approach could be used to generate simulated power demand time histories for input into a building energy simulation that permit analysis of the uncertainty surrounding the loads and future studies will examine the impact of such a stochastic load profile on the simulation output. The parameterisation chosen facilitates assimilation of monitored data as they become available, and also may be extended to different types of use. The study presented here has concentrated on building zones occupied by students, administrative staff and classrooms as those are the data available to us in the format required. It is not unreasonable to suggest that the schedules identified here for the administrative staff would be applicable to many office users, and indeed a consideration of the National Calculation Methodology activity database (Building Research Establishment, 2016) indicates that the schedules assumed for office workers are similar across different industries. A similar approach could be used to analyse lighting power demand for user-controlled lighting making allowances for seasonal performance. It also seems viable that the simulated diversity profiles might be used as a proxy for occupancy diversity. This is particularly useful as occupancy monitoring is not widespread; even where occupancy is recorded accessing such data is not straightforward.

Conclusions

This paper presents an efficient way to parameterise plug loads in buildings for input into a building energy simulation and to quantify the uncertainty in those loads. We have introduced an alternative parameterisation in which the plug loads are characterised not just by peak demand and diversity, but by peak demand, daily load range and a diversity factor on the load range which varies from 0 to 1 daily. This benefits a building energy simulation by not restricting the base load to be a fixed proportion of the peak daily demand.

Having specified the parameterisation in this way, we have then extracted diversity data from monitored plug loads for four zones in the same building occupied by different types of user. Few studies have been published to date which assess the variability in diversity between building zones or use types, or which attempt to accurately reflect the timing of the daily peak demand. Yet data analysis reveals differences between different use types that may be leveraged to improve simulation. In the Functional Data Analysis approach used here the diversity data have been separated into amplitude and phase functions followed by functional Principal Component Analysis of each element separately. The variability in the shape of the diversity curve has then been quantified for the different use types using a reduced number of functional principal components to keep complexity to a minimum. The results show good agreement with monitored data for both the zones analysed and a test zone in the same building.

Future studies will investigate the application of this approach to different end-uses and to different use types across a more diverse building portfolio incorporating the commercial sector. It is our intention to incorporate the uncertainty in occupant-related loads into a building energy simulation and to assess the corresponding impact on the simulation output. In this way the wider applicability of the modelling approach will be explored.

Acknowledgements

This study is supported by EPSRC I-Case PhD Studentship with Laing O'Rourke and by EPSRC grant (EP/L024452/1): Bayesian Building Energy Management (B-bem).

References

Building Research Establishment (2016, November). National Calculation Methodology (NCM) modelling guide (for buildings other than dwellings in England). Technical report.

de Wilde, P. (2014, May). The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction 41*, 40–49.

D'Oca, S. and T. Hong (2015, February). Occupancy schedules learning process through a data mining framework. *Energy and Buildings 88*, 395–408.

Feng, X. et al. (2015, January). Simulation of occupancy in buildings. *Energy and Buildings 87*, 348–359.

Khosrowpour, A. et al. (2016, September). Occupant workstation level energy-use prediction in commercial buildings: Developing and assessing a new method to enable targeted energy efficiency programs. *Energy and Buildings 127*, 1133–1145.

Liang, X. et al. (2016, June). Occupancy data analytics and prediction: A case study. *Building and Environment 102*, 179–192.

Mahdavi, A. et al. (2016, October). Prediction of plug loads in office buildings: Simplified and probabilistic methods. *Energy and Buildings 129*, 322–329.

Mahdavi, A. and F. Tahmasebi (2015, January). Predicting people's presence in buildings: An empirically based model performance analysis. *Energy and Buildings 86*, 349–355.

Marron, J. S. et al. (2015, November). Functional Data Analysis of Amplitude and Phase Variation. *Statistical Science 30*(4), 468–484.

Menezes, A. et al. (2014, June). Estimating the energy consumption and power demand of small power equipment in office buildings. *Energy and Buildings 75*, 199–209.

Rice, A. et al. (2010). A limited-data model of building energy consumption. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, pp. 67–72. ACM.

Rysanek, A. M. and R. Choudhary (2015, March). DELORES an open-source tool for stochastic prediction of occupant services demand. *Journal of Building Performance Simulation 8*(2), 97–118.

Srivastava, A. and E. Klassen (2016). *Functional and Shape Data Analysis.* New York: Springer.

Sun, Y. (2014). *Closing the building energy performance gap by improving our predictions.* Ph. D. thesis, Georgia Institute of Technology.

Tucker, J. D. (2015, December). fdaSRVF_matlab. `https://github.com/jdtuck/fdasrvf_MATLAB`.

Tucker, J. D. et al. (2013, May). Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis 61*, 50–66.

Wang, Q. et al. (2016, July). Meta-modeling of occupancy variables and analysis of their impact on energy outcomes of office buildings. *Applied Energy 174*, 166–180.

Ward, R. et al. (2016a, September). Data-driven bottom-up approach for modelling internal loads in building energy simulation using functional principal components. Newcastle, UK. BSO16, IBPSA England.

Ward, R. et al. (2016b, July). Exploring the impact of different parameterisations of occupant-related internal loads in building energy simulation. *Energy and Buildings 123*, 92–105.

Wilkins, C. and M. H. Hosni (2000). Heat gain from office equipment. *Ashrae Journal 42*(6), 33.

Yan, D. et al. (2015, November). Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings 107*, 264–278.