MEASURED INTERNAL TEMPERATURES IN UK HOMES - A TIME SERIES ANALYSIS AND ARIMA MODELLING APPROACH

Argyris Oraiopoulos[1, 2], Tom Kane[1, 2], Steven K Firth[1, 2], Kevin J Lomas[1, 2]
[1]Building Energy Research Group (BERG),
School of Civil and Building Engineering, Loughborough University, Loughborough, UK
[2]London-Loughborough (LoLo) Centre for Doctoral Training on Energy Demand,
Loughborough University, Loughborough, UK

ABSTRACT

This paper presents an analysis of internal air temperatures measured hourly in the living rooms of 10 domestic buildings in the city of Leicester, UK. Time series analysis is used to develop empirical models of room temperatures in rooms that are neither mechanically heated nor cooled, during the summertime period of July and August 2009. The models are used in predicting future temperatures based on past measured values. Such models can enable overheating risk alerts for homeowners and public authorities to be more accurately estimated and targeted.

INTRODUCTION

As global temperatures rise and the climate becomes more unstable, heatwaves will be a more common phenomenon (HPA, 2012). This could result in an increase of energy consumption in UK homes during summer periods due to a higher demand for cooling, but it could also have a substantial impact on heat related morbidity and mortality rates and produce a series of challenges for the emergency services and the national health system (Grogan & Hopkins, 2002). Overheating risk in domestic buildings is often predicted using modelling techniques based on assumptions of heat gains, heat losses and heat storage (Hacker et al., 2005, Porritt et al., 2012). Often dynamic thermal simulation software is used in which the modeler is required to decide a number of input assumptions upon which the result is depended. These assumptions often lead to modelling errors and reduce confidence in the results. Recent large-scale data collection studies allow empirical approaches based on measurements alone. Such methods could base the prediction of internal temperatures in dwellings, on previously recorded internal temperatures and external climate data.

Time series analysis has been successfully used in fields such as economics, geophysics, control engineering and meteorology to describe, explain, predict and control processes (Chatfield, 1996). Time series data are not simply data collected over time; there has to exist some form of ordering. A definition is given by Bloomfield (1976), "*A collection of numerical observations arranged in natural order with each observation associated with a particular instant of time or interval of the time which provides the ordering would qualify as time series data*". The analysis of time series data can be done either in the time domain, where the data are described in terms of the statistical relationships between observations at different times or in the frequency domain, where the fluctuations in one or more series are described in terms of sinusoidal behavior at various frequencies (Reddy, 2011).

The aim of the research presented in this paper is to apply the time series analysis method in the field of building physics and more specifically to room temperature data. The study is based on time domain analysis and within that there are three approaches to modelling the behaviour of a series; the smoothing methods, the ordinary least squares models and the stochastic models. Stochastic or probability models are those that are used to calculate the possibility of a future value lying between two specified limits (Box & Jenkins, 1970). Such models have been used to predict thermal loads in homes (Pfafferott et al., 2005, Ogunsola et al., 2014) and thermal conditions in hospital wards (Iddon et al., 2015). This novel approach is used to explore the mechanisms of the formation of such data series and to develop statistical models that allow the prediction of future temperatures based on past measured values and external climate data.

The application of these statistical models, could lead to the provision of tailored advice to occupants on how and when to act in order to reduce indoor temperatures during hot summer conditions. It could also allow timely information to those caring for the elderly and infirmed in order to prevent adverse health impacts due to increased temperatures in enclosed spaces. By applying an empirical predictive model to national datasets, it can provide significant insights for the developments of future policies in mitigating overheating in homes across the country and allow for a more detailed plan to be issued in the event of a heatwave. Finally, with the aid of the latest developments in generating future external weather data for the 2030s, 2050s and 2080s (Eames et al., 2011), at-risk households can be supplied with information on how to reduce the risk of overheating in the future.

METHODOLOGY

Household survey

The data used in this study were collected in Leicester during the summer months of 2009 as part of the 4M project (Lomas et al., 2010), which focused on representing carbon emissions from different sources to measure the carbon footprint of the city of Leicester. One of the project themes was *Building Energy*, which investigated the energy demand of the city's domestic buildings. A face-to-face questionnaire was administered to 575 houses that documented the house type, the house age, the type of wall (solid, cavity, filled cavity) and the number of occupants.

The largest proportion of the houses was semi-detached (41.7%) with mid-terraces covering more than a quarter of the sample (27%), together accounting for almost 70% of the sample. Concerning the age of the houses, 20% were built before 1920, 31.3% between 1920 and 1944 and 30.9% after 1965. Almost 44% of the houses have solid walls, while 53% have more than 200mm of loft insulation. More than a third of the sample has two occupants, with the vast majority being above 30 years of age at the time of the survey in 2009.

Temperature data collection

Hobo pendant type temperature sensors were used to record internal temperatures in the living rooms and main bedrooms over an eight-month period, starting on 1 July 2009. The sensors recorded air temperature at hourly intervals, however as they were not shielded, they will also have recorded a radiant component. From the 951 Hobo sensors that were deployed in 481 houses (94 households did not agree in taking the sensors) only 416 were found to contain valid data from free-running homes (no heating or cooling present) for the same 62 day period between 1st July and 31st August 2009. From the 416 sensors, in 230 homes, 212 are from living rooms and 204 are from bedrooms, hence some of the houses have only got measured temperatures from a single room. This data have already served as a solid basis for research projects focusing on indoor temperatures both in the summer (Lomas and Kane, 2013) and winter (Kane et al., 2015). An example of measured temperature profiles in a living room and a bedroom is given in Figure 1. The external weather data were obtained from De Montfort University, in the middle of Leicester, for the purposes of the 4M project. Figure 2 illustrates the external temperature measured hourly, together with the external mean temperature and the solar irradiation data as recorded during the period between the 1st July and 31st August 2009. It can be observed that the monitoring period started with some very high temperatures, while the lowest temperature, 7.9°C, was recorded in the middle of the monitoring period.

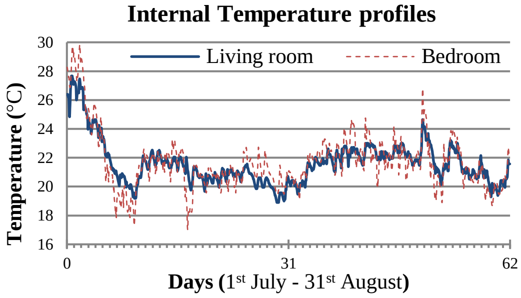


Figure 1 Hourly measured room air temperatures of example home between 1st July and 31st August 2009

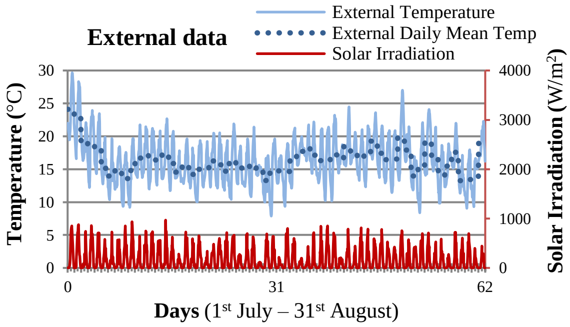


Figure 2 External weather data measured hourly between 1st July and 31st August obtained from De Montfort University

Time series analysis

The objectives of the analysis are based on the univariate time series modelling construction theory derived from the work of Box and Jenkins (1970), outlined below in Figure 3.

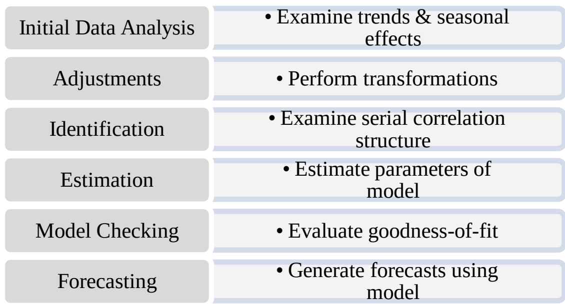


Figure 3 Time series analysis objectives, (after Box and Jenkins, 1970)

The first step in time series analysis is to describe the data by plotting and obtaining simple descriptive measures of the main properties of the series, checking for trends and seasonal variations as well as for outliers and unusual observations. Secondly one needs to perform suitable transformations to ensure that the series is converted to stationary (a series is considered to be stationary if it exhibits no trend (no systematic change in the mean) and no other seasonal or cyclical variations (strictly periodic variations)). Then, by examining the structure of the sample autocorrelation function (the plot of the correlation coefficient between a measurement and one *k* hours (lags) apart (otherwise known as correlogram or the autocorrelation function (ACF)) and the partial autocorrelation function (PACF), one aims to determine the form of the time series model that could provide the best fit to the data. Once the parameters of the best-fit model are estimated, the forecast ability of the model is evaluated by examining different statistical criteria, such as the coefficient of determination (R^2), the Root Mean Squared Error (RMSE) and the Bayesian Information Criterion (BIC). Also by using a proportion of the data to develop the model, one can then use the remaining of the measured data to check the forecasted values against them.

ARIMA models

In a statistical analysis of a time series, the ARIMA (Auto Regressive Integrated Moving Average) models provide a parsimonious description of a non-stationary stochastic process. The ARIMA model construction is a general linear framework that consists of three sub models: the autoregressive (AR), the moving average (MA) and the integrated (I). The first two components address the stochastic component of the series, while the integrated component is responsible for converting the series to stationary. More specifically, the autoregressive part of the model captures the past behaviour of the series, whereas the moving average explains the random shocks on the system (Reddy, 2011). ARIMA models are denoted by ARIMA (*p, d, q*) where *p* is the order of the regular autoregressive part, *q* is the order of the regular moving average part and *d* is the number of times the series had to be differenced in order to be converted to stationary. Before giving a general notation of an ARIMA model it is essential to describe an operator that is a useful notation when working with time series and ARIMA notations. This is the backward shift operator *B*, which for a time series Y_t given below.

$$B^d Y_t = Y_{t-d} \qquad (1)$$

Equation 1 means that when *B* is acting on Y_t it has the effect of shifting the data back by *d* time periods. A general notation of an ARIMA model is given by equation 2. For a time series Y_t , if the *d*-th order of differencing of the series is given by:

$$X_t = (1 - B)^d Y_t \qquad (2)$$

Then, the ARIMA (p, d, q) is given by Equation 3 below:

$$X_t = \varphi_t X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t - \theta_t \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \qquad (3)$$

Where φ_p are the estimated parameters for the autoregressive part of the model, $\theta_q$are the estimated parameters for the moving average part and ε_t is the errors at time t. In practise, numerous time series exhibit some form of a seasonal (periodic) component, which repeats after a specific amount of observations (*s*). Box and Jenkins (1970) have developed a general multiplicative seasonal ARIMA (SARIMA) model that deals with seasonality. Such models are denoted by ARIMA (*p, d, q*) $(P, D, Q)_s$ where *P* is the order of the seasonal autoregressive part, *Q* is the order of the seasonal moving average part and *D* is the seasonal differencing.

A general notation of a ARIMA (*p, d, q*) $(P, D, Q)_s$ is given below:

$$\left(1 - \varphi_1 B - \cdots - \varphi_p B\right)\left(1 - \Phi_1 B^s - \cdots - \Phi_P B^{s\times P}\right) (1 - B)^d (1 - B^s)^D Y_t = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right) \left(1 + \Theta_1 B^s + \cdots + \Theta_Q B^{s\times Q}\right)\varepsilon_t \qquad (4)$$

Where φ_p and Φ_P are the estimated parameters for the regular and seasonal autoregressive part of the model, θ_q and $\Theta_Q$are the estimated parameters for the regular and seasonal moving average part, *B* is the backshift operator given by Equation 1, ε_t are the errors at time t and *s* the seasonal period. Generally, SARIMA models assume that the model holding for one season holds for every other season as well. It is important to note that this equation results in a model that uses the previous values of a series to predict the value 1 time step ahead.

Initial study of a sub sample

Before proceeding with the analysis of the whole dataset, it is sensible to test this approach on a smaller, sub sample of the 230 homes. For this reason, 10 homes were selected for this analysis, based on the statistics of the household survey. Table 1 below outlines the characteristics of the selected homes.

Table 1 Characteristics of sub-sample

Houses	*House type*	*House age*
House 1	Semi detached	Pre 1900
House 2	Semi detached	1920-1944
House 3	Semi detached	1920-1944
House 4	Semi detached	1945-1964
House 5	Mid-terrace	1900-1919
House 6	Mid-terrace	1920-1944
House 7	Mid-terrace	1945-1964
House 8	End-terrace	1965-1980
House 9	Detached	1965-1980
House 10	Flat	Post 1980

Houses 1-10 were primarily selected based on their house type, to be representative of the dataset. Therefore, 40% are semi-detached, 30% are mid-terraces and end-terraces, detached and flats are all 10% each. Regarding their age, 30% of the sub sample houses were built between 1920 and 1944, 30% after 1965 while 20% were built before 1920. The temperatures measured in the living rooms of these houses will form the dataset for the analysis. Following is the method used to identify the best model fit for the data.

Model identification

The first stage in model identification is to produce the ACF and PACF of the data. If the series has positive autocorrelations up to an increased number of lags then most likely it needs at least one order of differencing to be converted to stationary. Once the series has been converted to stationary then by examining the new structure of the ACF and PACF is likely to determine one or more possible models. However, in practise the process of identifying the best-fit model is a lot more complex since the spikes in the structure of both the ACF and PACF many times are not as profound and therefore a number of models need to be considered, tested and evaluated using a number of model fit criteria before deciding on the most appropriate. This process is assisted by computer software in order to reduce its time length. In this analysis the model identification is done with the aid of the software SPSS.

Model fit criteria

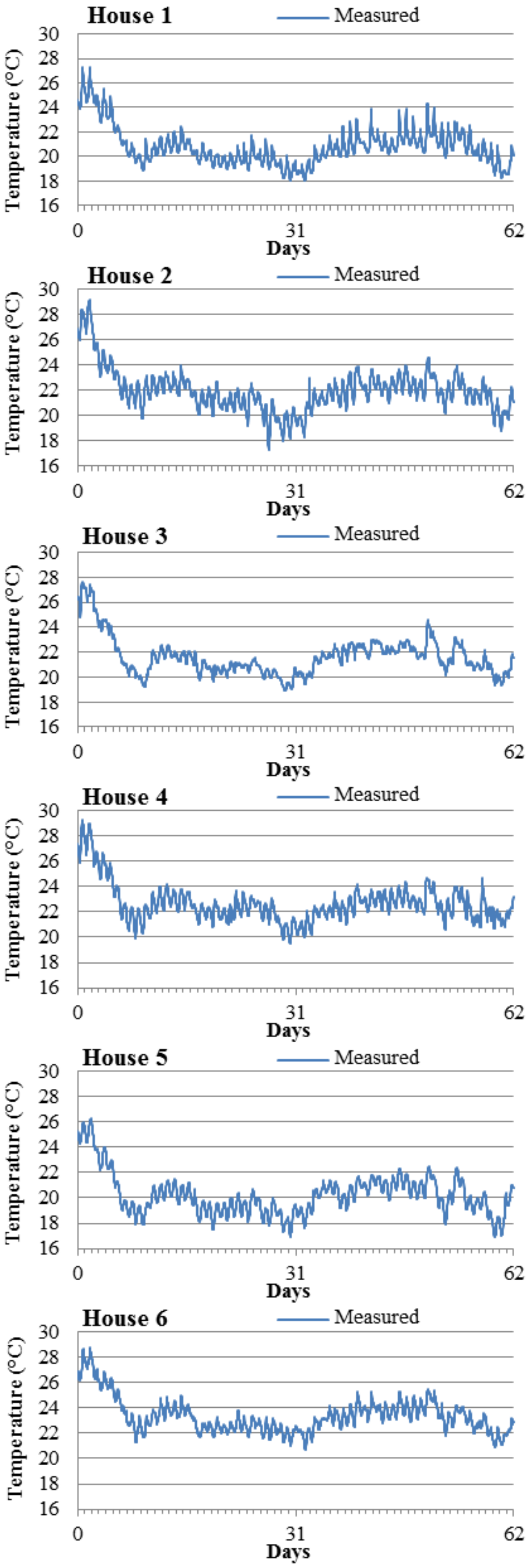
This analysis is making use of three different criteria in evaluating the computed ARIMA models; the coefficient of determination, the root mean square error and the Bayesian information criterion. The coefficient of determination (R^2) is an estimate of the proportion of the total variation in the series that is explained by the model. The root mean square error (RMSE) is a measure of the differences between the values predicted by the model and those actually measured during the survey. The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models where the model with the lowest BIC value is chosen. This is the only criterion of the three that attempts to account for model complexity by including a penalty for the number of parameters in the model and therefore reduces the likelihood of model overfitting. The penalty removes the advantage of models with more parameters, making the statistic easy to compare across different models for the same series. Whereas the coefficient of determination is a measure that allows for inferences regarding the goodness of fit of a single model, the root mean square error and the Bayesian information criterion are measures that can only be used when comparing models. Below are the results for the modelling of the sub sample of the 10 houses.

RESULTS AND DISCUSSION

Here are first presented the measured temperatures in the living rooms of the sub sample of 10 houses that have been selected for this analysis. Following are the results of the Univariate ARIMA model development. Initially the structure of the models is identified, then the statistics of the models are computed and finally the forecasts are compared with the measured values. It is important to note that the development of the Univariate ARIMA models was based on the data measured during the first 60 days of the monitoring period and the data measured during the final 2 days were used to make forecasts and compared them against the measured data.

Analysis of measured Temperatures

The following figure presents the hourly measured internal temperature in the living rooms of the 10 houses between the 1st July and 31st August 2009 in Leicester, UK.



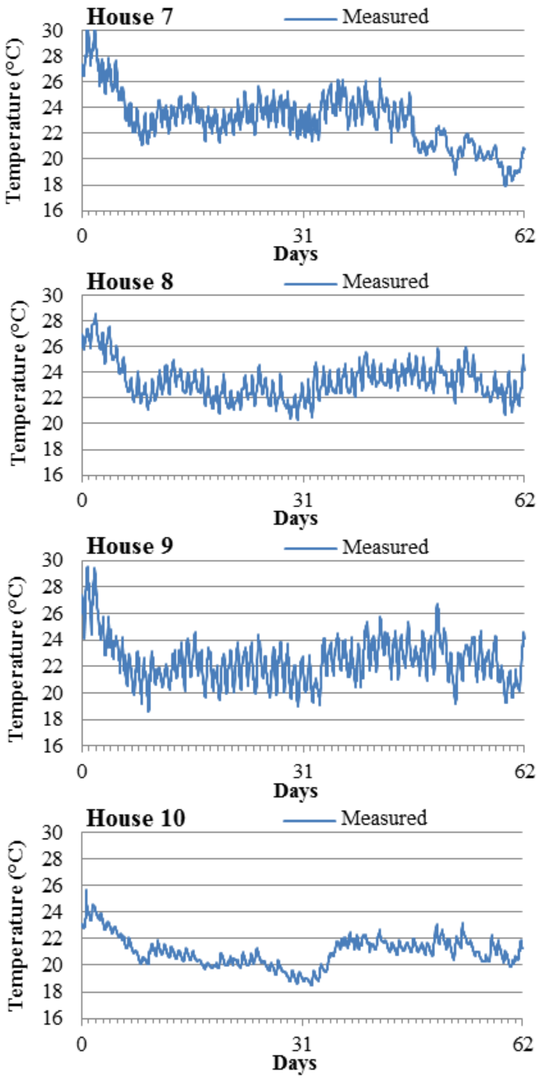


Figure 4 Hourly room air temperatures measured in living rooms of 10 sub sample houses between 1st July and 31st August 2009

From observing the individual graphs it is evident that the temperature series for all living rooms in the 10 houses are non-stationary (they do exhibit periodic variations) and therefore the series should be differenced before attempting to identify the most appropriate model structure. Furthermore it should be noted that some houses present a larger temperature variation than others, both throughout the monitoring period of the 62 days and within individual days. For example houses 7 and 9 present a temperature range of 12 °C throughout the 62-day monitoring period while the temperatures in houses 6, 8 and 10 ranged up to 8 °C. Also houses 7 and 9 present a larger day to day variation in relation to houses 3, 5 and 10. The reason for that amongst others could be the infiltration rates of the houses, the orientation of each space or the occupants' behaviour. In the context of overheating, the orientation of the houses (or the specific spaces) is an important parameter the effects of which require defined solutions, such as overhangs, blinds or low-e coating in glazed surfaces, however the behaviour of the occupants' is of paramount importance as particular actions can affect the indoor environment and therefore increase the risk of overheating.

Model Structure

As mentioned in the methodology section, the identification of the best-fit model is done with the aid of the forecasting commands of SPSS. By employing the expert modeller and selecting the ARIMA modelling technique, the software computes the most appropriate model for each space and estimates its parameters. By making use of the first 60 days of the monitoring period the models developed for the 10 houses are outlined in the following table.

Table 2 Structures and statistics of Univariate ARIMA models fit for the first 60 days of the monitoring period

Houses	*R^2*	*RMSE*	*BIC*	*(p, d, q) (P, D, Q)*
House 1	0.983	0.188	-3.318	(1,1,2) (0,1,1)
House 2	0.989	0.176	-3.455	(1,1,2) (1,0,1)
House 3	0.992	0.141	-3.908	(1,1,0) (1,0,1)
House 4	0.987	0.175	-3.463	(1,1,6) (1,0,1)
House 5	0.995	0.104	-4.508	(1,1,1) (0,1,1)
House 6	0.987	0.155	-3.720	(0,1,0) (1,0,1)
House 7	0.975	0.303	-2.360	(0,1,3) (1,1,1)
House 8	0.986	0.166	-3.575	(2,1,0) (1,0,1)
House 9	0.962	0.326	-2.215	(1,1,2) (1,0,1)
House 10	0.989	0.117	-4.271	(1,1,0) (1,0,1)

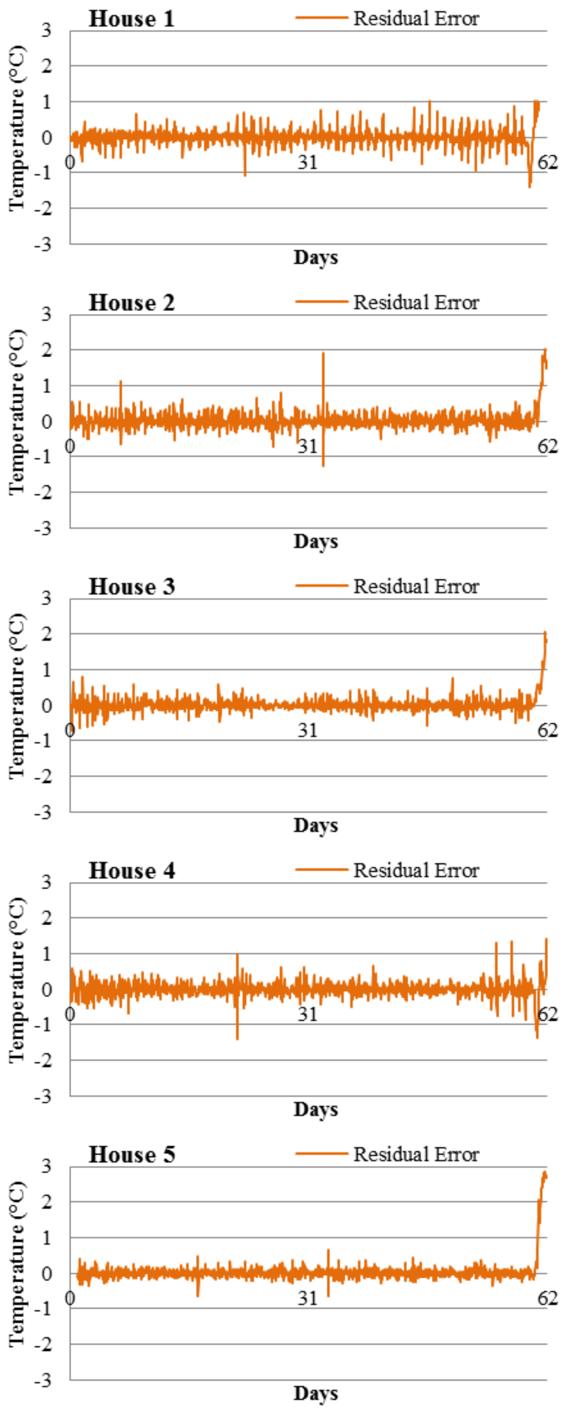
The above table shows the models that were fitted for the first 60 days of the monitoring period as well as the statistics associated with them. Having obtained the estimated parameters φ_p, Φ_P, θ_q and Θ_Q from the software output, by using Equation 4, the model equation for House 1 is given below:

$$(1-0.81Y_{t-1})(1-Y_{t-1})(1-Y_{t-24})Y_t = (1+0.634B+0.241Y_{t-2})(1+0.867Y_{t-24})\varepsilon_t \quad (5)$$

Similarly the equations for the rest of the models can be formulated. All the series have a regular difference of first order and in addition models for houses 1, 5 and 7 also have a seasonal differencing component. This means that the software has identified a stronger seasonal pattern in these series than in the rest of the houses, which have identical seasonal components. The statistics of the models are based on the fitted values of the model for each 1-hour time step. Every 1-hour time step the model predicts the value of the temperature 1-hour ahead using the past measured values, then an R^2 value and an error is calculated for this prediction and these are all summed up and averaged at the end of the fit period of 60 days. It is clear that there is a relationship between the three criteria for each space but as outlined in the methodology, each one of the criteria denotes a different capacity of the models. By comparing the values of the criteria and the measured data (Figure 4), it is apparent that the series that present the largest temperature swings, have both larger RMSE and BIC (houses 7 and 9) in comparison to the rest of the houses.

Residuals

Following are the results of the residual errors for the models fitted for the first 60 days of the monitoring period, for each house.



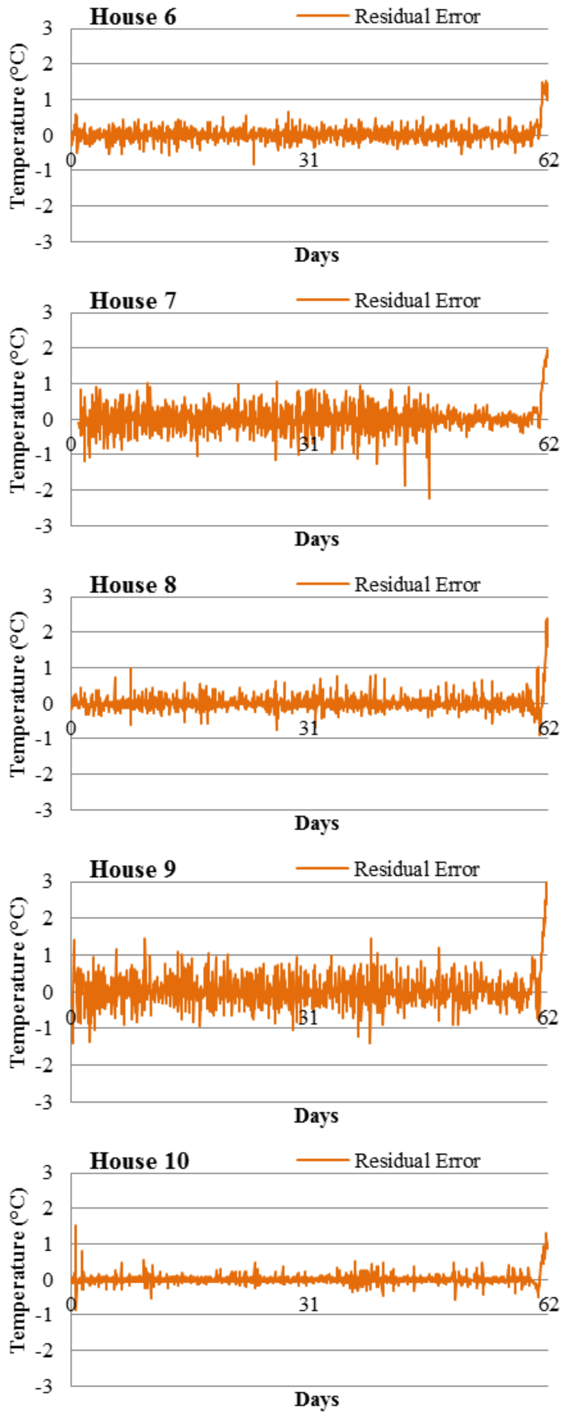
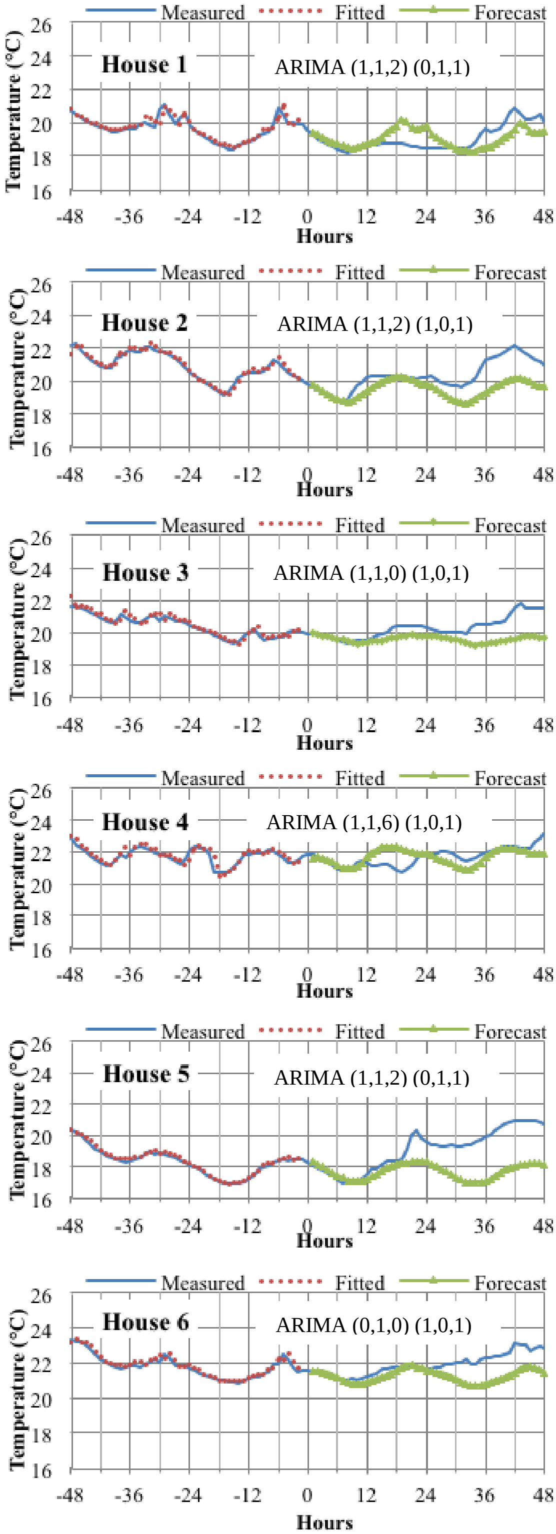


Figure 5 Residual errors between measured and fitted values of models for each house the10 sub sample houses between 1st July and 31st August 2009

The graphs in Figure 5 above illustrate the residual errors of the developed Univariate ARIMA models. It has to be noted that the high values at the right end of each graph are due to the fact that the first 60 days are included in the development of the models while the last 2 days are just used for the forecasts. This will become clearer in the following section. Overall, the patterns of the residuals indicate that the houses with the least smoothed measured temperature profiles (houses 7 and 9) present the largest errors even when the best-fit model is calculated.

Forecasts

This section presents the results of the 2-day forecast for each house. The graphs in the following figure illustrate the measured data together the fitted values of the models for and the forecasted values for the last 2 days of the monitoring period.



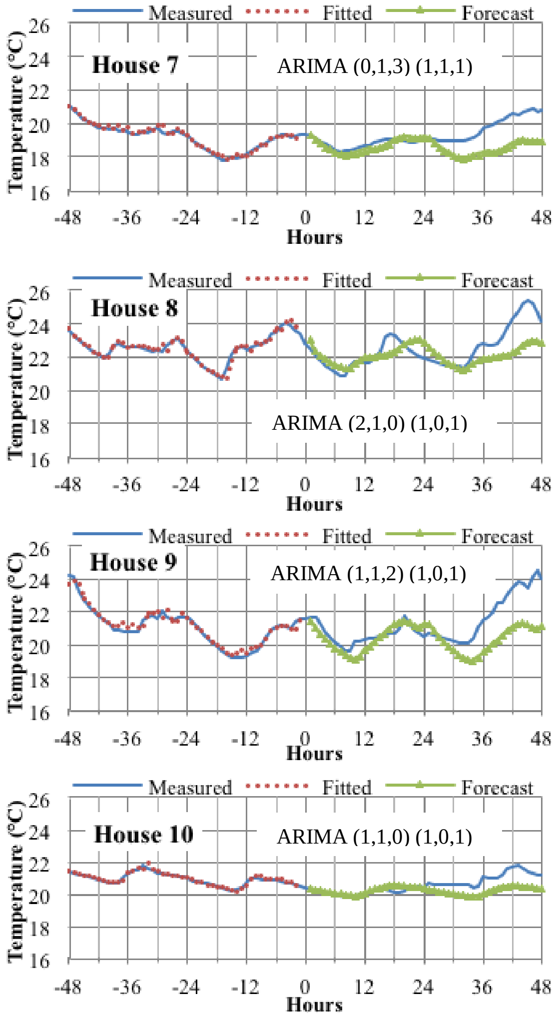


Figure 6 Measured, fitted (each time step is 1 hour ahead prediction based on past measurements) and forecasts (not based on measured values but only on previous predicted values) for sub sample

The graphs in the figure above illustrate the differences between the measured, the fitted and the forecasted values of the Univariate ARIMA models. In the fitted series, the value at every time step is based on the previous measurements, while the forecasted series that starts at the end of the 60-day period for which the models were developed (0 hours), is based on previously predicted values and not measured. All the models can predict quite accurately the temperatures for the first 6 hours and some for the first 12 or even 24. However it is clear that after the first 24 hours the errors presented are fairly high with most of the models exhibiting differences of up to 2 °C at the end of the 2-day (48 hours) forecast period. It is also interesting that models that presented small residual errors and would be assessed as better performing than others according to the model statistics (House 5), present big differences between the measured and the predicted values. Since the Univariate ARIMA models developed are based only on previous values of the measured data, it is sensible to conclude that the inclusion of the external weather data as independent variable in the models structure could potentially to improve the predictability of this modelling approach.

CONCLUSION

This paper presents the time series analysis of the internal air temperatures in the living rooms of 10 houses, measured between 1st July and 31st August 2009 (62 days) in Leicester, UK. The first 60 days of the observed data have been used to develop statistical Univariate ARIMA models, which have been assessed using 3 different statistical criteria and the last 2 days of the monitoring period have been used to produce forecasts. The extent of the measured internal temperature swing presents a close relationship with both the model fit criteria and the residuals between the measured data and the fitted values of the models. Finally, the forecasted values for the last 2 days of the monitoring period reveal the reduced ability of the ARIMA models to predict temperatures up to 48 hours ahead. There are 4 main conclusions drawn from this study:

- Univariate ARIMA models can be used to model the internal temperatures in houses based on past measurements but only up to 12 hours ahead.
- The increase of the measured internal temperature swing increases the residual error of the models.
- The extent of the residual errors and the goodness of fit of the model do not relate to the ability of the model to forecast future values accurately.
- It is essential to include the external weather data and develop Multivariate statistical models to improve the time length of the predictability of the models by responding more accurately to the changes of the external temperature.

Such models could predict future internal temperatures based on past values and by including the external temperature, they could provide essential information regarding overheating alerts during hot summer conditions.

ACKNOWLEDGEMENT

This research was made possible by the support from the Engineering and Physical Sciences Research Council (EPSRC) for the London-Loughborough Centre for Doctoral Research in Energy Demand (grant EP/H009612/1). The 4M consortium was funded by the EPSRC under their Sustainable Urban Environment programme (grant EP/F007604/1).

REFERENCES

Bloomfield, P. 1976 *Fourier analysis of time series: an introduction*. New York: John Wiley.

Box, G.E.P., Jenkin.s, G.M. 1970. *Time series analysis: forecasting and control*. San Francisco: Holden-Day.

Chatfield, C. 1996. *The analysis of time series: an introduction*.5th ed. London: Chapman & Hall.

Eames, M., Kershaw, T., Coley, D. 2011. On the creation of future probabilistic design weather years from UKCP09. *Building Services Engineering Research and Technology*; 32:127-42.

Grogan, H., Hopkins, P.M. 2002 Heat stroke: implications for critical care and anaesthesia. *British Journal of Anaesthesia*; 88: 700–07.

Hacker, J.N., Belcher, S.E., Connell, R..K., 2005 Beating the Heat: keeping UK buildings cool in a warming climate. UKCIP Briefing Report. UKCIP. Oxford, UK.

Health Protection Agency (HPA) 2012. Health Effects of Climate Change in the UK 2012: current evidence, recommendations and research gaps. London, UK.

Iddon, C.R., Mills, T.C., Giridharan, R., Lomas, K.J. 2015. The influence of hospital ward design on resilience to heat waves: an exploration using distributed lag models. *Energy and Buildings*; 86:573-88.

Kane, T., Firth S.K., Lomas, K.J., 2015. How are UK homes heated? A city-wide, socio-technical survey and implications for energy modelling. *Energy and Buildings*; 86:817-832.

Lomas, K.J., Bell, M.C., Firth, S.K., Gaston, K.J., Goodman, P., Leake JR, Namdeo A, Rylatt M, Allinson D, Davies, Z.G., Edmondson, J.L., Galatioto, F., Brake, J.A., Guo, L., Fill, G., Irvine, K.N., Taylor, S.C., Tiwary, A. 2010. The carbon footprint of UK cities: 4M: measurement, modelling, mapping and management, ISOCARP Review 06, International Society of City and Regional Planners. p.168–91.

Lomas, K.J., Kane, T. 2013. Summertime temperatures and thermal comfort in UK homes. *Building Research & Information*; 41:259-80.

Ogunsola, O.T., Song, L., Wang, G. 2014. Development and validation of a time series model for real time thermal load estimation. *Energy and Buildings*;76:440-49.

Pfafferott, J., Herkel, S., Wapler, J. 2005. Thermal building behaviour in summer: long term data evaluation using simplified models. *Energy and Buildings*; 37:844-52.

Porritt, S.M., Cropper, P., Shao, L.C., Goodier, C.I. 2012. Ranking of interventions to reduce dwelling overheating during heat waves. *Energy and Buildings*; 55:16-27.

Reddy, T.A. 2011. *Applied data analysis and modeling for energy engineers and scientists*. New York: Springer.