# IEA Annex 58 statistical guidelines for building energy data

In many years the wells of data from buildings have become richer and a great potential for extracting useful information has build up. Dealing with this data the challenges are manifold – in this short post we will touch on the statistical modelling techniques needed. In very general terms one can say, that the basic challenge is to select the best model fitting the observed data, however in doing that, some steps must be considered:

- How to define a model
- How to estimate the parameters in a model
- How to measure the fit of a model and use this to select the best model
- How to extract the information which is useful for the particular application

Indeed, ultimatively an applied model should be evaluated on the usefulness of the information it provides for the application it is used for, however that is often not practically possible. Therefore procedures leading to the selection of a suitable model must be formed – statistics provides the needed techniques. Observed data varies: some of the variation is systematic and some of the variation is random. The statistical techniques enable a mathematical understanding of the systematic relation between observed variables, as well as a mathematical understanding of the random variation. They can be used, both to learn about informative parameters in the models, e.g. the heat loss coefficient of a building, as well as finding models suitable for prediction and control. Most statistical methods can be derived with the principle of maximum likelihood, which provides a clear and transparent methodology for defining models, estimating parameters and measuring the fit of a model to data.

Time series data from building energy systems exhibits naturally strong dynamical relations stemming from the thermal processes of the systems. Luckily, the dynamics can often successfully be described by models for time invariant linear (LTI) systems, on the relevant time resolutions (e.g. hourly values). Statistical models for LTI systems are very well understood, see (Box et al., 1976) and Madsen (2007). The discrete models for LTI systems are referred to as Auto-Regressive Moving Average with eXogenous input, in short ARMAX, models. Several software implementations exists, here the Matlab system identification toolbox (Ljung, 2008), and the R inbuilt function ARIMA and the MARIMA package Spliid (2017), are recommended. Also useful for modelling of LTI systems are continuous time models, where the dynamical relations between the variables are described by differential equations, hence the model and its parameters have a direct meaning in the context of physics. To facilitate the method of maximum likelihood for continuous time models, thus enabling the entire statistical methodology, the continuous time models must be formulated as stochastic differential equations – hence a proper description of the random variation is included and the models are called a grey-box models. A very useful implementation for grey-box modelling is available in the R package ctsmr (Juhl et al., 2016). Further, it is perfectly possible to include time-varying and other non-linear effects in both discrete and grey-box models, even in a non-parametric way – hence the exact functional relationship is not defined, only some constraints of its shape. This can be carried out for example with base spline or Fourier functions, even while keeping some parts of the model parametric, thus forming a semi-parametric model.

In order to provide statistical modelling procedures, which can be used to extract information about the thermal performance of buildings and their systems, quite some work was carried out in IEA Annex 58. One of the documents, referred to as the IEA Annex 58 statistical guidelines (Madsen et al., 2015), holds the description a statistical procedure developed to model thermal performance of buildings using simple data, e.g. hourly values of internal and external temperature, heat input and solar radiation. The statistical guidelines provide a procedure for selecting a suitable ARX model and from the selected model to calculate the heat loss coefficient and gA-value. Further, the basics of grey-box modelling for building energy systems are introduced. The guidelines can be downloaded together with examples in R from [1]. The statistical guidelines is a good base for the statistical models and techniques, which will be needed in Annex 71. Time-varying and non-linear phenomena, e.g. caused by occupants, must be modelled in an effective way, e.g. as in (Bacher et al., 2013) where a diurnal curve is included in a discrete model using Fourier series.

[1] http://orbit.dtu.dk/en/publications/thermal-performance-characterization-using-time-series-data--iea-ebc-annex-58-gu .html

## References


Peder Bacher, Henrik Madsen, Henrik Aalborg Nielsen, and Bengt Perers. Short-term heat load forecasting for single family houses. *Energy and Buildings*, 65(0):101–112, 2013. ISSN 0378-7788. doi: http://dx.doi.org/10.1016/j.enbuild.2013.04.022. URL `http://www.sciencedirect.com/science/article/pii/S0378778813002752`.

G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis*. Holden-day San Francisco, 1976.

R. Juhl, N. R. Kristensen, and H. Madsen. *Continuous Time Stochastic Modeling in R*. DTU Compute, 2016. URL `ctsm.info`.

Lennart Ljung. *System Identification Toolbox 7: Getting Started Guide*. The MathWorks, 2008.

Henrik Madsen. *Time Series Analysis*. Chapman & Hall/CRC, 2007.

Henrik Madsen, Peder Bacher, Geert Bauwens, An-Heleen Deconinck, Glenn Reynders, Staf Roels, Eline Himpe, and Guillaume Lethe. *Thermal Performance Characterization using Time Series Data - IEA EBC Annex 58 Guidelines*. Technical University of Denmark (DTU), 2015.

Henrik Spliid. *marima: Multivariate ARIMA and ARIMA-X Analysis*, 2017. URL `https://CRAN.R-project.org/package=marima`. R package version 2.2.