CLIMA 2016 - proceedings of the 12th REHVA World Congress

Heiselberg, Per Kvols

Publication date:
2016

Document Version
Final published version

Data-Driven Modelling of the Energy Use in Dwellings using Smart Meter Data

Eline Himpe[1], Arnold Janssens[2]

Department of Architecture and Urban Planning, Ghent University
Sint-Pietersnieuwstraat 41-B4, 9000 Ghent, Belgium

[1]eline.himpe@ugent.be

[2]arnold.janssens@ugent.be

Abstract

The increasing application of energy monitoring and smart metering systems leads to an increasing availability of long-term measurements of the actual energy use in occupied residential buildings. By use of data-driven modelling techniques, the actual energy use of buildings can be modelled for different purposes, such as normalisation to the outdoor climate, forecasting, parameter identification or decomposition. Heating degree day and energy signature methods are widely applied data-driven energy models, that are typically used when occasional measurements are available over long periods of time. However with the availability of high frequency energy monitoring data, the applicability of these classical methods is questioned from statistical point of view. On the other hand, more accurate models and more extensive information retrieval are expected from the analysis of these data.

In this paper various data-driven modelling techniques are applied on a data-set that includes hourly gas and weather monitoring data of 25 dwellings. Starting from linear regression models on weekly and daily energy use data during a heating season, the time step and length of the data-set are gradually reduced and the use of linear input-output models and time series decomposition methods are explored and illustrated. It is found that auto-regressive models with exogenous inputs are useful alternatives to the linear regression models when the time step of measurements is reduced to daily and 2-hourly data. Furthermore, time series decomposition enable the separation of deterministic diurnal patterns from long-term data trends in 2-hourly datasets.

Keywords – energy monitoring; smart meter; residential buildings; normalisation

1. Introduction

The growing interest in the energy performance of buildings brings along questions about the real energy performance of occupied buildings. While building simulations are useful to estimate the building performance during the design phase, these theoretically estimated energy use figures often differ significantly from the real energy use of the occupied building.

Reasons for this can be, for example, assumptions used in the simulation tools, building construction deficiencies, the actual operation and control of the building services, occupant behavioural aspects and the interaction between the building, the services and the occupants [1]. On the other hand, the increasing application of energy monitoring and smart metering systems, leads to an increasing availability of frequent and long-term measurements of the actual energy use in occupied buildings. By use of data-driven modelling techniques the energy use of buildings can be modelled, starting from actual energy use figures, measurements of weather variables and possibly other energy-related parameters (e.g. indoor temperature) [2]. Dependent on the application in view (e.g. energy feedback, identification of energy savings), the models can be used to normalise the data with respect to weather conditions, to decompose different types of energy use in the data, to predict future energy use or to estimate energy use characteristics.

Heating Degree Day and Energy Signature models are widely applied data-driven energy models, that are typically used when occasional measurements are available over long periods of time. However with the availability of energy monitoring data that are more frequent (e.g. measurement time steps between 15 minutes and 1 hour), the applicability of these classical methods is questioned from statistical point of view [3], and on the other hand more accurate results and more extensive information retrieval are expected from the analysis of these data. In this paper different data-driven modelling techniques are applied on a data-set that includes hourly gas and weather monitoring data for 25 Belgian dwellings that use gas only for space heating. The models are compared in terms of statistical quality, model fit and one-step-ahead prediction accuracy for different levels of aggregation and length of the data-set.

The next section of this paper presents the case-study data. In section 3 the data is aggregated to daily values and modelled using linear regression models (LR) and auto-regressive models with exogenous inputs (ARX). The length of the measurements is reduced from an entire heating season to 1 month and the impact of the weather variables on the needed data length is discussed. In section 4, the time-step of the data points is reduced to 2 hours and the use of auto-regressive models with exogenous inputs (ARX), and time series decomposition methods (TSD) are explored and illustrated.

2. Case study data

The case-study data includes hourly gas and weather monitoring data for 25 Belgian dwellings that use gas only for space heating (not for other energy use functions) and use only gas for space heating (no other energy carriers). The data-set reaches from February 2011 to April 2012. The dwellings are located in the same village, but they are different in typology, age, building constructions and services and users. One dwelling (n° 3) is irregularly occupied and often unoccupied during the measurement period.

3. Daily aggregated data

A. *Weekly to daily aggregated data*

Heating degree day and energy signature methods use linear regression methods to fit an energy balance equation (1), in which E is the energy (gas) use of the building, I_s is the global horizontal solar irradiation and T_a is the ambient temperature measured at a national weather station. c_1, c_2 and c_3 are the model coefficients and $\boldsymbol{\varepsilon}$ is the error term. Both heating degree days or outdoor temperature can be used as an input, but the use of equivalent temperatures typically leads to better model fit. Also the use of solar irradiation as an input yields an equal and often better model fit for all of the studied houses (see Fig. 1 (right)).

$$E = c_1 + c_2 \times HDDeq + c_3 \times Is + \varepsilon \qquad (1)$$

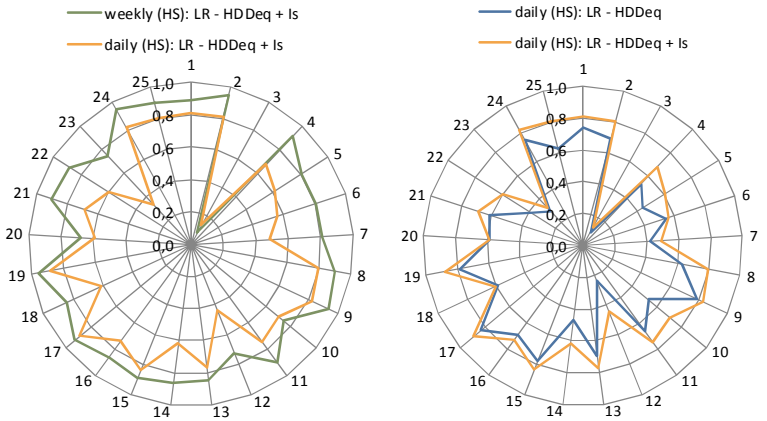

Fig. 1: R^2adjusted for Linear Regression models for 25 dwellings:
(left) using weekly vs. daily time steps, (right) with vs. without solar radiation I_s as an input

The model described in (1) is applied on data from one heating season of the case-study dwelling, and the use of weekly and daily data points is compared in Fig. 1 (left). For all 25 case study houses, the coefficient of determination R^2_{adj} is clearly lower for daily data then for weekly data. A reason for this is that in the daily aggregated data, the data points are not independent. They are correlated in time because of physical phenomena such as the thermal inertia and behavioural patterns (e.g. week days vs. week-end days) with periods bigger than one day (and the use of equivalent temperatures is resolving this issue only partially). In weekly aggregated data, these phenomena are actually averaged out so the subsequent data points are independent. An auto-regressive model with exogenous inputs

(ARX) is now applied to the daily aggregated data set. In this kind of model, historical values of the output E_t and eventually of the inputs ($T_{a,t}$ and $I_{s,t}$) are added as extra inputs in the model. The most simple example is the 1st order ARX model in equation (2), where the energy use of the previous day E_{t-1} is an input. Like linear regression models, ARX models can also be fitted using the ordinary least squares estimation. As can be seen in Fig. 2 (left), the ARX(1) model already leads to better model fits in all of the houses except for house 3, which is a house with very irregular occupation. The use of higher order models ARX(+) leads to a small improvements on the ARX(1) model in many of the case-study houses (Fig. 2 (left)).

$$E_t = c_1 + c_2 \times Ta_t + c_3 \times Is_t + c_4 \times E_{t-1} + \varepsilon \qquad (2)$$

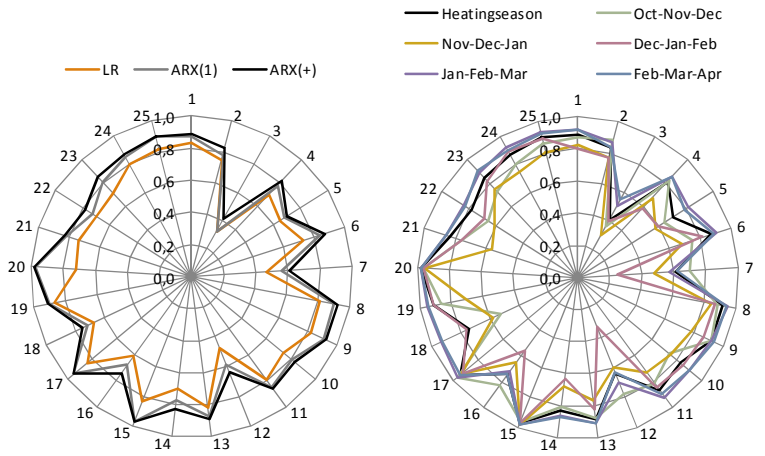

Fig 2 R^2adjusted for 25 dwellings: (left) LR and ARX models during 1 heating season (right) ARX(+) models during 1 heating season and 3 month data subsets

B. Length of the data-set and data selection

The ARX(+) model is now fitted to sub-datasets of the heating season with a length of 3 months, and with a length of one month. In Fig. 2 (right) the R^2_{adj} values obtained with a data length of 3 months are compared to those obtained with the entire heating season data-set for all 25 dwellings. It is found that for some time periods (e.g. January+February+March and February+March+April), the model fit is always equal to or better than when the entire heating season is used, and this is also the case for the RMSE, MAE and standard errors (Table 1). For other moments of the heating season (e.g. November+December+January) the model fit is clearly worse, or the

findings are different for different houses. When the data length is further reduced to one month, similar conclusions can be made.

	R^2adj		St. Error		RMSE		MAE	
Heating season	*0,83*	0,97	*18,6*	7,8	*18,2*	7,8	*14,2*	5,0
Jan-Feb-Mar	*0,90*	0,98	*18,1*	8,9	*17,3*	8,4	*13,8*	5,6

Table 1 Statistics for *Linear Model* and ARX(1) model

Whether the model is intended for use to estimate the model coefficients or to normalise or predict the space heating energy use, it is important that the dataset that is used for fitting the model includes the variety of weather conditions that can be expected. When taking a look at the plots of ambient temperature and solar irradiation in Fig. 3, it is clear that for the considered heating season, the ambient temperatures are very similar from November to end of January, and during this period the solar irradiation is rather low. Therefore a model fitted on the data-set November+December+January will be less suitable to distinguish the effects of ambient temperature from those of solar radiation, or to predict space heating energy use in a periods with more sun. On the other hand, the period January+February+March covers a large variety in both temperatures and solar irradiation, and moreover the cross-correlation between these two weather inputs is rather low.

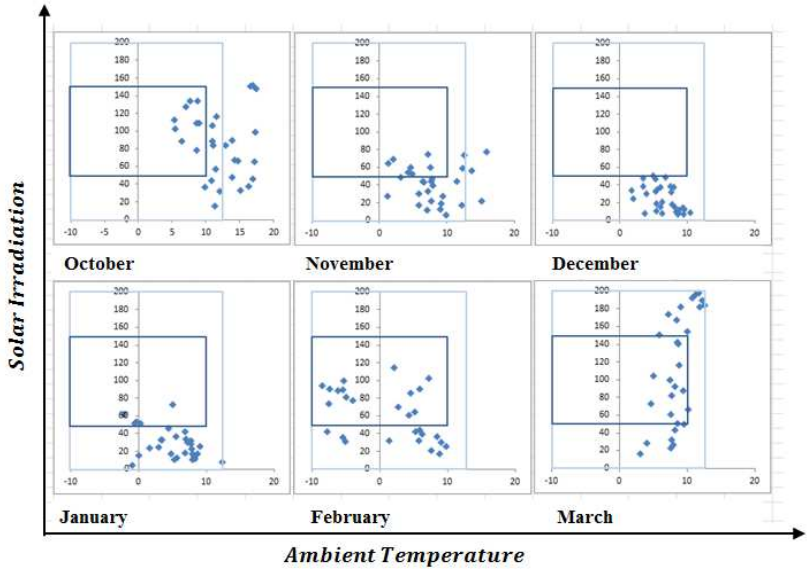

Fig. 3 Daily Solar Irradiation and Ambient Temperature for each month in winter 2011-2012

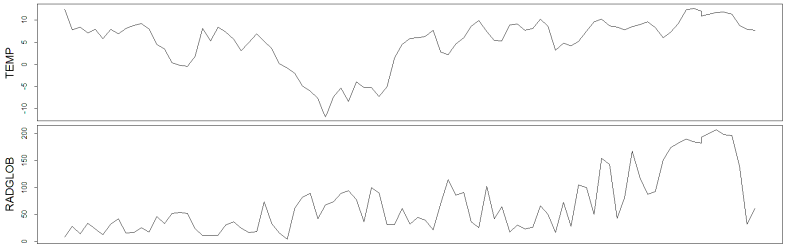

Fig. 4 Time series of ambient temperature and solar irradiation for January+February+March

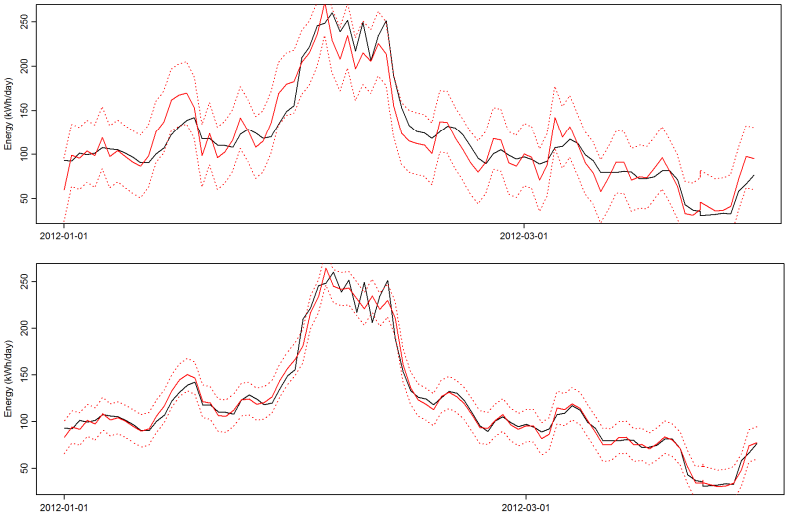

Fig. 5 Time series of observed and fitted values with 95% confidence bounds for case-study house 15 for January+February+March, by use of: (top) LR and (bottom) ARX(1) models

From observations of the daily Belgian data for different years empirical criteria for the weather variation are roughly estimated (remark that these criteria are indicative, they should be further validated in future work):

(1) $T_{a,min}$ < ca. 3°C; ΔT_a > 15°C;

(2) $I_{s,max}$ > 120 W/m² and ΔI_s > 70W/m² and

(3) Cross-Correlation $T_a \sim I_s$ < 0,50.

It is concluded that the length of the dataset can be reduced to relatively short periods between 1 and 3 months, while maintaining a good model fit and one-step ahead predictions, on the condition that the above criteria for the weather inputs are met. Fig. 4 shows the observed values and the in-sample one-step ahead predictions with 95% confidence intervals for the linear regression model and the ARX(1)-model for a data length of three months from beginning of January to end of March 2012.

4. 2-hourly data

Smart metering systems typically record the energy meter readings with time steps between 15 minutes and 1 hour. In this section, the data are aggregated to 2-hourly values and possibilities for modelling these data are explored, using a one month dataset (January 2012) of case dwelling 15.

A. ARX models

When the time-step of the data is smaller than one day, diurnal variations in weather inputs and the energy use, that are due to the heating system control and occupant behaviour, are noticeable (see Fig. 7). As can be expected from the findings in section 3, ordinary linear regression models (LR) are not suitable to model this kind of data: the statistical model assumptions are not fulfilled, the model fit is bad and the one-step ahead predictions cannot follow the fluctuations in the data (Fig. 7 and Table 2).

January	**R^2adj**	**St. Error**	**RMSE**	**MAE**
LR	0,01	20,2	10,2	6,5
ARX	0,99	2,4	1,2	0,9
TSD	-	-	1,4	1,1

Table 2 Statistics for models on 2 hourly data

Again ARX models can be fitted to the data. Because of the diurnal variation in the data, it is now necessary not only to include the energy use of the previous time step, but also of 12 time steps or 24 hours ago, as an important input in the model described in equation (3).

$$E_t = c_1 + c_2 \times Ta_t + c_3 \times Is_t + c_4 \times E_{t-1} + c_5 \times E_{t-12} + \cdots + \varepsilon \quad (3)$$

This input proved to be common for the data-sets of almost all of the 25 dwellings that were modelled. Then, dependent on the specificities of the data-set, extra inputs can be added to take into account the remaining correlations in the data: typical inputs are temperatures or solar radiation of the previous hours or days, or the energy use one week ago (in case a weekly pattern in the data occurs). Statistical indicators of the common model for the exemplary dwelling are given in Table 2, and indicate a good model quality. Also the plots of the one-step-ahead predictions in Fig. 7 present a much better fit.

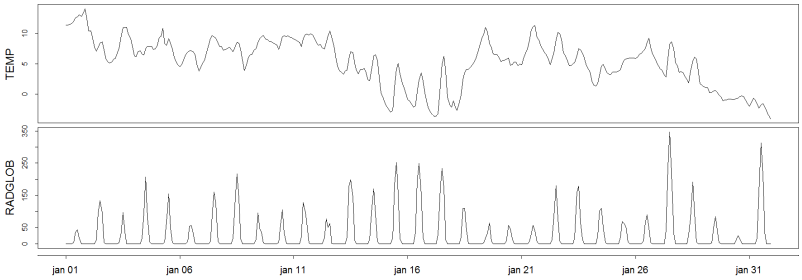

Fig. 6 Time series of ambient temperature and solar irradiation for January 2012

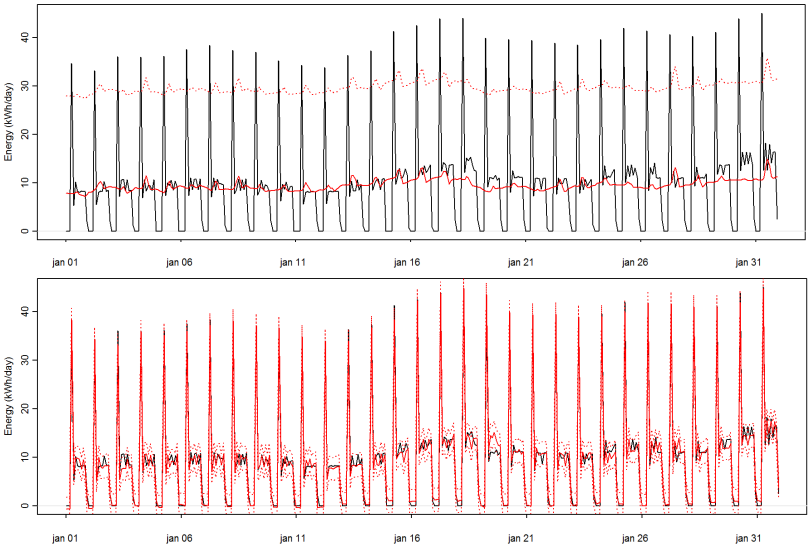

Fig. 7 Time series of observed and fitted values with 95% confidence bounds for case-study house 15 for January, by use of: (top) LR and (bottom) ARX models

B. Time series decomposition

When observing the time series plots of the energy use (Fig. 6), two patterns are easily recognised, that is a pattern related to the diurnal variation, and a pattern related to the long-term behaviour of the series (e.g. due to weather influences). While the ARX-models (Fig. 7 (bottom)) are able to capture the combined effect of these patterns, they do not characterise them in a deterministic way.

Therefore another approach to deal with the 2-hourly data-set is to decompose the time series of the energy use into the diurnal and the long-term pattern, while irregularities remain in the error component. In the field of time series analysis, the diurnal pattern in this data-set would be called a 'seasonal component', that is a component that is influenced by seasonal factors (in casu: the time of the day) and has a fixed period (in casu: 24 hours). The long-term variation in the data is then called the 'trend component'. Various approaches exist to estimate the time series components. In this work a classical additive decomposition is applied: by use of moving averages, the trend component is estimated, and after the data is de-trended, the seasonal component is calculated by averaging the data over all the periods. Finally the error term is calculated by subtracting the trend and seasonal terms from the observed data.

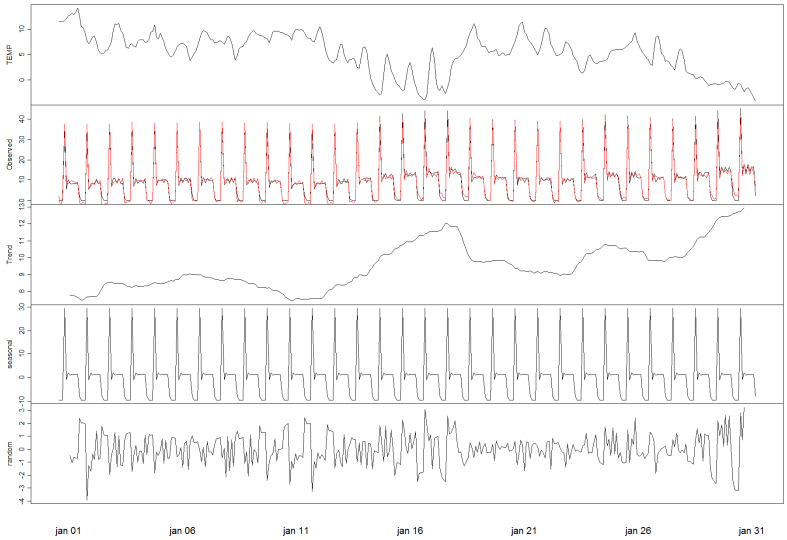

Fig. 8: Time series decomposition of the energy use data of a dwelling. Plot of the outdoor temperature, observed energy use data, trend component, seasonal component and the residuals

The time series decomposition (TSD) is applied to the case-study dwelling and presented in Fig. 8. The seasonal component is a pattern with a fixed shape which, in this case, is clearly related to the control of the heating system: at night a set-back is applied and the heating turns off, while the restart of the heating in the morning causes a peak in gas use. The trend represents the slowly increasing and decreasing energy use over the course of the days and weeks, and, when comparing plots in Fig. 8, for this house, it is clearly related to the variations in ambient temperature. However, remark that ambient temperature itself was not an input when estimating the trend,

and, in cases where other long-term phenomena influence the energy use of the house, these will also be influencing the trend. Therefore the analysis of the different components of the energy use time series could be useful to identify properties of or changes in energy use patterns (e.g. due to changes in the heating system or occupancy profile) as well as long-term changes (e.g. due to energy-efficiency measures or behavioural change).

In Fig. 8 also the predicted values are plotted for the TSD method and the accuracy statistics are calculated (Table 2). Both the RMSE and MAE for this model are in the same order of magnitude as in the ARX-model presented in section 4A, indicating that for the in-sample predictions, both models are more or less as accurate.

5. Conclusion & Perspectives

In this paper data-driven modelling techniques are applied on datasets containing 2-hourly and daily aggregated natural gas and weather smart metering data of 25 dwellings. It is found that because of auto-correlation in the data, the goodness-of-fit of classic linear regression models are lower when the time step of the data is reduced from weekly to daily or 2-hourly values. Then auto-regressive models with exogenous inputs are a straight-forward alternative. Regarding the needed length of the dataset, it is concluded that data lengths of one up to three months can result in the same model quality as when an entire heating season is used, on the condition that the input weather data provide sufficient variation. Furthermore it is found that time series decomposition methods can provide similar accuracy in one-step-ahead predictions as the ARX-models, but they also enable the separation of deterministic diurnal patterns from long-term data trends in the 2-hourly datasets. From this exploration, it is found that ARX-models can be useful when the objective is to describe the energy use and estimate model coefficients, while both ARX and TSD models can be used to make energy use predictions with a short prediction horizon. However for the application of long-term predictions or normalisation of the data, the applicability of these models should be further investigated.

Acknowledgement

The authors are grateful to *Eandis*, the Flemish Distribution System Operator, for providing access to the monitoring data of their smart metering proof-of-concept project.

References

[1] M. Delghust. Improving the predictive power of simplified residential space heating demand models: a field data and model driven study. Ghent University, Ghent, Belgium, 2015.
[2] M. Santamouris. Energy Performance of Residential Buildings – a practical guide for energy rating and efficiency. London, 2009, Earthscan.
[3] S. Hammarsten. A critical appraisal of energy-signature models. Applied Energy 26 (1987) pp. 97-110.