Comparing statistical modeling techniques for heat loss coefficient estimation using in-situ data

Xiang Zhang*, Katia Rito, Dirk Saelens and Staf Roels
KU Leuven, Department of Civil Engineering, Building Physics and Sustainable Design, Kasteelpark Arenberg 40 Bus 2447, 3001 Heverlee, Belgium

Abstract. The combination of in-situ collected data and statistical modelling techniques proved to be a promising approach in actual building energy performance assessments, such as heat loss coefficient (HLC) evaluation. In this study, based on datasets of co-heating and pseudo-random binary sequence heating tests on a portable site office, the performance of three types of statistical models (i.e. multiple linear regression (MLR), autoregressive with exogenous terms (ARX), and grey-box models) on HLC-determination are examined. It is revealed that a similar HLC estimation outcome (about 115 W/K) is offered by the aforesaid three types of statistical models, but with different confidence intervals (CIs), where the 95% CIs of MLR (±3.1%) and ARX (±2.4%) are relatively narrow and the ones of grey box models are somewhat wider (around ±9%). Moreover, for the current case study building, with evenly orientation-wise distributed glazed envelope, integrating B-splines in to the grey-box model, to characterize the solar aperture (gA) and solar gain dynamics more precisely, imposed insignificant effects on the HLC estimation and corresponding 95% CIs, compared to the grey-box model with a constant gA assumption.

Keywords. Heat loss coefficient, solar gain, solar aperture (gA), grey-box model, autoregressive with exogenous terms (ARX) model.

1. Introduction
The heat loss coefficient (HLC), in W/K, is a crucial indicator to reveal the building energy performance, which is widely applied in building construction practice assessment as well as in the selection of building refurbishment actions or for heating system design [1]. It is defined as the heat flow rate per unit temperature difference (W/K) between building interior and exterior, or the power (W) required to maintain a 1K temperature difference over the building fabric [2]. Thus, the HLC, is a joint indicator for quantifying both thermal insulation and airtightness performance of the building envelope [3]. To account for the performance gap between designed and actual achieved thermal performance, based on on-site measured data, statistical modelling approaches have been put forward to estimate the actual buildings’ HLC. Those approaches range from simple (e.g. linear regression) to more complex methods - autoregressive with exogenous terms (ARX) and grey-box modelling. It is interesting to compare and explore the HLC estimation performance of those statistical methods in particular applications. Besides, solar gains are another vital physical aspect estimated by those methods. Thus, it is also interesting to assess the impact of the elaborated solar gain dynamics modeling on the HLC-determination. Rasmussen et. al. [4] has proposed a dynamic solar aperture (gA) integrated grey-box modelling technique to capture dynamic solar gains, which is subsequently

*corresponding author: jason.zhang@kuleuven.be; zhangxiang@cqu.edu.cn (X. Zhang)
studied as the elaborated solar gain modelling approach in this study. Hence, based on on-site measured datasets from a portable site office (PSO) case (see section 3.1), in this study, the effects to HLC estimations from 1) three different statistical models and 2) integration of dynamic solar aperture (gA) in grey-box modelling (i.e. comparing to the constant gA integrated one) are examined.

2. Statistical modelling approaches

2.1. Multiple linear regression

Treating a building as one zone, the simplified indoor energy heat balance can be expressed as Equation (1) [5]. In Equation (1), heating input (Ph) is the dependent variable, while indoor to outdoor temperature difference ΔT (i.e. Ti – Te) and global horizontal irradiance (Isol) are the independent ones. This multiple linear regression technique, in three dimensions, allows determining both heat loss coefficient (HLC) and solar aperture (gA).

\[ Ph = HLC \times (Ti - Te) - gA \times Isol \]  

(1)

2.2. ARX model

The ARX model is a linear representation of a dynamic system in discrete time, which is a standard methodology in handling a linear time-invariant (LTI) system. In the abbreviation, AR represents AutoRegressive and X indicates that eXternal inputs are considered, denoting that the assumption of the ARX model predicts the system output based on not only the previous system output but also previous (or current) additional system inputs. The most common ARX-model structure is a linear differential equation that connects the current output at time t to a finite number of past outputs and inputs [6], shown in Equation (2):

\[ A(q^{-1})y(t) = B(q^{-1})u(t) + e(t) \]  

(2)

Where u(t) is the input, y(t) is the output and e(t) is the equation error while A(q⁻¹) and B(q⁻¹) are polynomials in the backward shift operator q⁻¹ [7], elaborated in Equation (3-4):

\[ A(q^{-1}) = 1 + a_1 q^{-1} + a_2 q^{-2} + \cdots + a_n q^{-n} \]  

(3)

\[ B(q^{-1}) = b_0 + b_1 q^{-1} + b_2 q^{-2} + \cdots + b_n q^{-n} \]  

(4)

The order (n) of polynomials generally reveals how many data observations from the past (also called lags) are modelled. Modelling the thermal behaviour of a building, given a 1st order example, the indoor temperature Tİₑ are at time t can be modelled as the output as a function of the external inputs at time t, such as the exterior temperature, heating input, and solar gain (Teₑ, Qi, and Sg₁), and one lag previous indoor temperature Tİₑ₋₁. See section 4 for more details.

2.3. Grey-box model

Integrating the advantages of both white-box models (fully physically based) and black-box models (fully data-based), grey-box models have been developed to combine the prior physical knowledge with statistical information embedded in the data [8,9]. The physical knowledge contributes to model buildings by analogy with electric circuits, mathematically represented as a system of stochastic first-order differential equations (SDEs) [10], making the equations and parameters involved physically interpretable. Statistical tools such as maximum likelihood (ML) are commonly adopted for parameter estimation[11]. For example, the simplest one-state RC-network model (denoted as Ti model) consists of only an indoor air temperature state Ti-equation (the reason called Ti model) and the measurement equation, Equation (5-6).
\[ dT_i = \left( \frac{1}{R_{ia}C_i} (T_e - T_i) + \frac{1}{C_i} (gA I_{sol} + \phi_h) \right) d_t + \sigma_i d\omega_i \]  

(5)

\[ Y_k = T_{ik} + e_k \]  

(6)

3. Study case and datasets

3.1. Portable site office (PSO)

In this study, the observed building is a portable site office (PSO) (Figure 1) which represents a simplified building. The PSO is located at a car park (50.86N, 4.68E) of the Arenberg campus, KU Leuven, Belgium. The main façade with the entrance is southeast-oriented (140° from the north) (Figure 3). The envelope of the PSO has a total of five windows, visible in Figure 3, of which four of them have the same size (with glazing part of 0.98*0.72m²). While the last one on the northwest wall, with a transparent area of 0.33*0.4m².

3.2. Reference value determination

As stated, the heat loss coefficient (HLC) is the key interest in this study. To determine the HLC reference value a co-heating test [12] is conducted from 19 December 2019 to 8 January 2020. As shown in Figure 4 as an example, bulbs (375 W each) are used as electrical heaters, and electrical fans were working as air mixers. The indoor temperature is homogeneously heated up to an elevated steady-state one (i.e. 25°C) [13]. The indoor temperature (\( T_i \) in °C) and heating input (\( Ph \) in W) (the energy delivered to maintain \( T_i \) as 25°C) were recorded in a five-minutes frequency. Meanwhile, corresponding exterior air temperature (\( T_e \) in °C) and global horizontal radiation (\( GHI \) in W/m²) were
measured in one-minute sampling time at a weather station at the VILET test building [14], located 300 m from the PSO, which can be considered as on-site. Thus, in total four aforesaid parameters (i.e. \( T_i, T_e, Ph, \) and \( GHI \)) are used for HLC estimation based on a multiple linear regression (MLR) model without a constant heat loss term (zero intercept), since a better regression result is generally obtained with a zero-intercept setting [5]. As illustrated in Equation (1), the HLC and \( gA \) are unknown coefficients in the MLR regression. For more details, we refer to [15]. It is noted that the measured five or one-minute(s) frequency data of the four parameters are averaged daily for the MLR application [13]. The reference HLC and \( gA \) were estimated to be 122.2 W/K and 1.97 m\(^2\), with multiple R-squared of 0.997. Note that, for the small size of the PSO (Figure 1-3), the HLC of 122.2 W/K is considerably high. This is mainly attributed to the significant heat loss through glass planes and window and door frames, as shown in the infrared image of Figure 2.

3.3. PRBS (pseudo random binary sequence) heating test

Dynamic models estimate parameters based on the dynamics (e.g. change features and patterns) of the measured data. In the other words, the richness of dynamics in the input signal is essential for a successful data fitting and parameters estimation. Hence, imposing a heating input uncorrelated with other signals (e.g. solar radiation and outdoor temperature) is vital. In this regard, a pseudo-random binary sequence (PRBS) signal is applied to the heating system (Figure 4), to switch on/off of the bulbs (heaters). To be specific, the key feature of PRBS signal [16] is the auto-correlation of white-noise properties and uncorrelated with other external signals (e.g. ambient weather conditions) [12,17]. Figure 5 gives an idea of the PRBS signal. Moreover, similar to the dataset construction in section 3.2, both ‘\( T_i \)’ and ‘\( Q \)’ are measured in 5 mins sampling time, and ‘\( T_e \)’, ‘\( GHI \)’ are recorded in one-minute frequency. To model the effect of wind, the wind speed (\( ws \)) is also monitored in the weather station. Finally, the aforesaid five parameters are averaged to hourly data and combined in one dataset for subsequent grey-box and ARX modelling in section 4. Figure 5 visualizes the constructed dataset.
4. Model selection

Both ARX and grey-box models have the theoretical endless potential to extend their model complexity, such as by increasing the number of orders in ARX and states in grey-box models. The statistical model with enhanced complexity generally has a better performance on data-fitting but along with an increased risk of ‘overfitting’. To identify a ‘suitable’ model complexity, in this study, the forward model selection, starting from the smallest feasible model and extending it step-by-step by adding one more variable at a time, is used [18]. Meanwhile, in general, a successful model extension is permitted under a considerable increase of the loglikelihood-value ($\text{LogL}$) and remarkable drops of AIC (Akaike information criterion) or BIC (Bayesian information criterion) values [18]. Figure 6 and Figure 7 visualize the processes of the optimal model identification for ARX and grey-box models.

For the ARX model, shown in Figure 6, starting from the order 1 model, the model is extended step-by-step by involving more previous observations (more lags) and removing insignificant parameters (marked as green in Figure 6) later on. Specifically, the ARX models in this study modelled indoor temperature ($T_i$) as the dependent variable, while the ones in the brackets (Figure 6) refer to the considered independent variables in the ARX model. Finally, the order 3 model in Figure 6 is selected as the optimal one, since the extension of the ARX model from order 3 to 4 didn’t lead to a remarkable drop in $\text{LogL}$, and even resulted in a series of statistically insignificant parameters (Figure 6).

The forward selection (i.e. embedding more components or states stepwise) also guides the optimal model selection in the grey-box models, and two identification steps are included (Figure 7). In step 1, the candidate grey-box models are based on a constant $gA$ assumption, meanwhile, the first-rate model identified in step 1 will serve the basis for the dynamic $gA$ integration in step 2. As shown in Figure 7, the TiTe_ws model is selected as the optimal one in step 1, owing to its lowest AIC (i.e. 553.27) and low $\text{LogL}$. Although the lowest $\text{LogL}$ (264.97) has been reached by the TiTe_Ria_ws model, this model was not selected as the $Ria$ parameter was found statistically insignificantly. Then, in step 2, B-splines are further integrated into the TiTe_ws model to reveal the dynamic $gA$. TiTe_ws_S4 is selected in step 2 (Figure 7) since very limited $\text{LogL}$ and AIC
improvements are observed by increasing the number of integrated B-splines (i.e. from S4 to S6 in Figure 7), and the BIC-value increased considerably during this process. Readers refer to [4] for the application details of B-spline customization and integration into the grey-box models. Figure 8 visualizes both the TiTe_ws and TiTe_ws_S4 models.

![Figure 7 Model selection of grey-box models](image)

![Figure 8 The selected two-states single zone thermal RC-network model of PSO](image)

5. Results and comparison

The four final statistical models mentioned, selected and for the PSO-building, are MLR, the ARX model of order 3 in Figure 6 (referred as ARX model in the following text), the grey-box model with constant gA (TiTe_ws model) and the dynamic gA based grey-box model (TiTe_ws_S4 model). As stated in section 1, the interests of this study are assessing the impacts of different models and different gA assumptions (i.e. constant and dynamic ones) on the HLC-determination. Thus, firstly, the estimated constant (or dynamic) gA with 95 percent confidence intervals (95CIs) is visualized in Figure 9. It is noticed that the variation on the estimated gA-value is significant. Considering that the total glass area of the PSO is 2.95 m² and assuming a g-value between 0.6-0.8, a gA-value around 2 m² seems reasonable. In this regard, MLR gives the closest estimation, but with a considerable wide 95CIs. The grey-box model significantly over-estimates the theoretically expected gA-value. One of the possible reasons is that the grey-box models might incorporate other physical phenomena into gA, such as the part of indirect solar gains transmitted through the opaque walls. For the current case, the
ARX model provides a reasonable (but slightly over-estimated) $gA$ assessment and relatively narrow 95CIs. Back to the main question of this study, Figure 10 shows that all three types of models (i.e. MLR, ARX, and grey-box models) give very similar $HLC$ estimates (about 115 W/K), with a relatively narrow 95CIs for the MLR and ARX-approaches (around $\pm3\%$), while the grey-box models, both constant or dynamic $gA$ based, show wider ones (up to $\pm9\%$ roughly), shown in Figure 10. Overall, the difference in $HLC$-determination performance of the four models is limited (Figure 10), and 95CIs of less than $\pm10\%$ in grey-box models can still be considered as a good performance. Interestingly, in this building with rather evenly distributed windows (Figure 3), integrating dynamic $gA$ into the grey-box model, to reach an elaborated solar gain estimation, imposes negligible effects on both the $HLC$ and corresponding 95CIs estimations.

![Figure 9 Estimated constant (or dynamic) solar aperture values with 95-percent confidence intervals](image)

![Figure 10 Estimated HLC values with 95-percent confidence intervals](image)

6. Conclusion

In this study, based on a simplified building (PSO), the performance of three types of statistical models (linear regression, ARX, and grey-box models) on $HLC$ estimation are explored. Taking the $HLC$-value, estimated by the MLR approach on co-heating test data, as a reference (i.e. 122.23 W/K), ARX and grey-box models provide slightly underestimated outcomes (around 111 W/K). While $HLC$ estimations performed with ARX and the two grey-box models (i.e. constant and dynamic $gA$ integrated ones) produce similar estimates, ARX could offer a slightly narrower 95CIs. Finally, in the PSO case with orientation-wise evenly distributed glazed envelope (Figure 3), integrating B-splines into the grey-box model, offering more precise dynamic information on the solar aperture and solar gain dynamics, imposed limited effect on $HLC$-estimate and its corresponding 95CIs.
References

[1] Senave M, Roels S, Reynders G, Verbeke S and Saelens D. Assessment of data analysis methods to identify the heat loss coefficient from on-board monitoring data. Energy Build 2020;209:109706.

[2] ISO. ISO 52016-1:2017. Energy Perform Build Needs Heat Cool Intern Temp Sensib Latent Heat Load—Part 1 2017.

[3] Senave M, Roels S, Verbeke S and Saelens D. Analysis of the influence of the definition of the interior dwelling temperature on the characterization of the heat loss coefficient via on-board monitoring. Energy Build 2020;215:109860.

[4] Rasmussen C, Frölke L, Bacher P, Madsen H and Rode C. Semi-parametric modelling of sun position dependent solar gain using B-splines in grey-box models. Sol Energy 2020;195:249–58.

[5] Bauwens G and Roels S. Co-heating test: A state-of-the-art. Energy Build 2014;82:163–72.

[6] Roels S, Bacher P, Bauwens G, Castaño S, Jiménez MJ and Madsen H. On site characterisation of the overall heat loss coefficient: Comparison of different assessment methods by a blind validation exercise on a round robin test box. Energy Build 2017.

[7] Diversi R and Guidorzi R, Soverini U. Identification of ARX and ARARX Models in the Presence of Input and Output Noises. Eur J Control 2010;16:242–55.

[8] Bacher P and Madsen H. Identifying suitable models for the heat dynamics of buildings. Energy Build 2011

[9] Deconinck AH and Roels S. Comparison of characterisation methods determining the thermal resistance of building components from onsite measurements. Energy Build 2016.

[10] Howell JR, Menguc MP and Siegel R. Thermal radiation heat transfer. CRC press; 2010.

[11] Kristensen NR, Madsen H and Jørgensen SB. Parameter estimation in stochastic grey-box models. Automatica 2004;40:225–37.

[12] Bauwens G. In Situ Testing of a Building’s Overall Heat Loss Coefficient - Embedding Quasi-stationary and Dynamic Tests in a Building Physical and Statistical Framework. KU Leuven, 2015.

[13] Arnold Janssens. Overview of Methods to Analyse Dynamic Data. International Energy Agency, EBC Annex 58, Reliable Building Energy Performance Characterisation Based on Full Scale Dynamic Measurements. Leuven: 2016.

[14] Roels S and Deurinck M. The effect of a reflective underlay on the global thermal behaviour of pitched roofs. Build Environ 2011;46:134–43.

[15] Roels S and Senave M. Determining the HLC from on site measurements: A building physical and statistical framework. IEA EBC Annex 71: Building energy performance assessment based on in-situ measurements. 2018.

[16] Godfrey KR. Correlation methods. Automatica 1980;16:527–34.

[17] Madsen H and Holst J. Estimation of continuous-time models for the heat dynamics of a building. Energy Build 1995;22:67–79.

[18] Claeskens G and Hjort NL. Model selection and model averaging. Cambridge Books 2008.