Building and Environment 44 (2009) 2378–2395

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

Interactions with window openings by office occupants

Frédéric Haldi*, Darren Robinson

Solar Energy and Building Physics Laboratory (LESO-PB), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 12 December 2008
Received in revised form
28 March 2009
Accepted 30 March 2009

Keywords:
Windows
Behavioural modelling
Building simulation

ABSTRACT

Based on almost seven years of continuous measurements, we have analysed in detail the influence of occupancy patterns, indoor temperature and outdoor climate parameters (temperature, wind speed and direction, relative humidity and rainfall) on window opening and closing behaviour. In this we have also considered the variability of behaviours between individuals. This paper begins by presenting some of the key findings from these analyses. We go on to develop and test several modelling approaches, including logistic probability distributions, Markov chains and continuous-time random processes. Based on detailed statistical analysis and cross-validation of each variant, we propose a hybrid of these techniques which models stochastic usage behaviour in a comprehensive and efficient way. We conclude by describing an algorithm for implementing this model in dynamic building simulation tools.

1. Introduction

1.1. State of the art

Pioneering investigations in residential buildings performed by Dick and Thomas in 1951 [1], Brundrett [2] from 1975 to 1979 and Lyberg in 1982 [3] reached agreement on the fact that actions on windows were positively correlated with external air temperature, and marginally negatively associated with wind speed. The influence of other stimuli was not examined.

In 1984, Warren and Perkins [4] showed – using stepwise multiple correlation analysis – that external air temperatures accounted for 76% of the observed variance, the sunshine for an additional 8% and wind speed for 4%. A linear relationship between the percentage of rooms with at least one window open (a Boolean response) was linked to external temperature. However this was not formally equivalent to a probability, as p may take values outside the range $[0, 1]$. A questionnaire conducted as a part of this study also revealed for the first time that occupants act on their windows particularly often on arrival and at departure.

A first attempt to develop a mathematical model to predict the state of windows was performed by Fritsch et al. in 1991 [5,6]. Based on measurements of the opening angle of four windows in four office rooms recorded every half an hour in the LESO building (see Section 2), a discrete-time Markov process model (see Section 3.2) was developed to predict transitions between bins of opening angles. The model is formulated as Markov chains defining transition probabilities between six states, each corresponding to a definite class of opening angles and adjusted for four different outdoor temperature ranges. The model includes transition matrices adapted to each occupant, in order to account for significant observed variations. The dependance of the percentage of opened windows versus wind speed and sunshine was also examined, but no significant variation was observed for wind speeds lower than 5–6 m/s. Although south-facing vertical irradiance was observed to be correlated, especially in the mid-season, only outdoor temperature was retained as a model parameter.

Towards the end of the 1990s, interest in the adaptive approach to thermal comfort drew attention to the relationship between behaviour and thermal satisfaction. This led to several measurement campaigns in Pakistan from 1993 to 1996 [8,9], the United Kingdom [10,11] and five European countries [12].

Based on measurements from these three surveys, Nicol [13] proposed in 2001 the first coherent probability distribution for the prediction of the state of windows, as logit functions (see Section 3.1) of indoor and outdoor temperature. In most cases the correlation with indoor temperature is similar to that with outdoor temperature, but Nicol recommends the use of outdoor temperature on the basis that it is an input of any simulation program, while indoor temperature is an output. However, Nicol and Humphreys [14] later reported in 2004 that indoor temperature was a more coherent predictor for the use of windows than outdoor temperature. This approach may seem more sensible: as Robinson [15] points out, predicted probabilities of interaction are otherwise independent of the design of the buildings in which occupants are accommodated.

* Corresponding author. Tel.: +41 21 693 55 54.
E-mail addresses: frederic.haldi@epfl.ch (F. Haldi), darren.robinson@epfl.ch (D. Robinson).

0360-1323/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.buildenv.2009.03.025

Rijal et al. [16,17] subsequently published in 2007 a more refined model, considering both indoor and outdoor temperature. A multiple logit distribution (with two variables) was derived for the probability of a window to be open. A deadband of ± 2 K for θ_{in} and ± 5 K for θ_{out} was defined to distinguish the probability of opening from that of closing. This refinement, although not based on any observed actual openings and closings, potentially solves the problem of repeated actions that would take place if a single distribution was used. The implementation in ESP-r of what has become named the *Humphreys algorithm*, involves the following steps:

1. Input parameters θ_{out}, $\theta_{out,rm}$, θ_{in} and θ_{op} are retrieved;
2. The comfort temperature θ_{comf} is computed from $\theta_{out,rm}$ according to the CEN standard:

$$\theta_{comf} = \begin{cases} 18.8 + 0.33\theta_{out,rm} & (\theta_{out,rm} > 10) \\ 22.6 + 0.09\theta_{out,rm} & (\theta_{out,rm} \le 10) \end{cases}$$

3. If $|\theta_{op} - \theta_{comf}| > 2K$, the probability of action is calculated by $\mathrm{logit}(p) = 0.171\theta_{op} + 0.166\theta_{out} - 6.4$;
4. This probability is compared with a random number to determine the next window state.

A comparison between observed and simulated window opening proportions for several indoor and outdoor temperature ranges is provided as validation.

Page developed in 2006 [18] a behavioural model of window openings, relating indoor temperature and pollutant concentration with interaction probability, based on thresholds defined by Fanger's thermal (ISO-7730) and aeraulic [19] comfort model. Interaction occurs when the calculated indoor concentration exceeds a critical concentration, or when hot or cold comfort limits (defined by $|\mathrm{PMV}| > 2$) are surpassed.

Based on their summer field survey, Haldi and Robinson [20] suggest that in summer the strong correlation between indoor and outdoor conditions in naturally ventilated buildings could dampen the efficiency of multiple logistic regressions. The works of Yun and Steemers [21,22] seem to strengthen this hypothesis. Rijal et al. [23] have subsequently published a refinement of the Humphreys algorithm, including a window opening effectiveness parameter. This modification imposes a window to be closed if $\theta_{out,rm} > 28.1$ °C and $\theta_{out} > \theta_{in} + 5$ °C.

Yun and Steemers [21,22] performed a field survey on 6 offices facing east, west or south in 2 buildings, during 3 months in summer only. Indoor temperature was retained as a driving stimulus, considering that *"the prediction as a function of external temperatures cannot be considered as an intrinsic result"*, in agreement with Robinson's observation [15]. It was noticed that changes in window states mainly occurred on arrival or at departure.

A useful feature of [21,22] is the use of separate probabilistic sub-models for the opening of windows at arrival and during occupancy. Retained offices did not enable night ventilation, so actions on departure are not considered (windows are assumed to be closed at departure). Furthermore, this model predicts changes in window state from open to closed and from closed to open using indoor temperature and previous window state as predictors. Analyses show that outdoor temperature is not significant, in agreement with the summer survey of Haldi and Robinson [20]. The probability of opening on arrival is a logit distribution of indoor temperature, while a linear function is proposed for actions during occupancy. A short comparison between participants is proposed. The final model, which has been implemented in ESP-r [35], retains thus indoor temperature, occupancy transitions and previous window state. Whilst this model may describe actions during the summer (for cases where deliberate night ventilation is not exercised), there is a question as to its validity during the winter period.

Herkel et al. [24,25] also pointed out that most window openings can be associated with the arrival of an occupant, and so proposed separate sub-models for window openings and closings on arrival, at departure and during occupancy. However, these sub-models consider outdoor temperature as the driving stimulus, based on the observation that this variable had a higher correlation with the hourly mean value of opening status of the monitored windows.[1] The final model is then formulated as six probabilities of opening and closing for arrival, intermediate presence and departure, given as quadratic functions of outdoor temperature.

Supported by two surveys (once in summer and once in winter), each consisting of a single questionnaire concerning window opening behaviour sent to 4948 dwellings in 2008, Andersen et al. [27] used multiple logistic regression analysis to deduce odd ratios for the significance of a set of variables. It was noticed that the respondent's gender, the outdoor temperature, the perceived illumination, air quality and noise levels had a statistically significant impact on "perceived" window opening behaviour.

1.2. Key advances, open questions and research needs

From the above review, we conclude the following:

- Explored methods for the simulation of window actions include logit distributions and discrete-time Markov processes.
- Thermal stimuli have been shown to be the predominant causes for actions (indeed non-thermal variables are generally ignored), but no clear consensus is reached whether indoor or outdoor temperature should be used as the independent variable in the simulation of actions on windows.
- It is known that the influence of occupancy patterns is important.
- Independent studies have observed specificities in summer behaviour, but seasonal variations in behaviour have yet to be taken into account.
- The treatment of occupant behaviour towards night ventilation is not considered.
- Window opening angles are mostly ignored, even though these are crucial for reliable air flow prediction.
- Published studies do not provide any common robust cross-validation procedure, which prevents any comparison of quality between published models.
- The case of offices with several occupants is not specifically treated (authoritarian versus democratic behaviour).
- Existing models are informed by measurements in office buildings and behaviour in residential environment is not specifically treated.

2. The field survey

In this section we present the experimental design that provided the basis for the development of our models, together with a description of the building in which the data were collected.

Data used here were collected from the Solar Energy and Building Physics Laboratory (LESO-PB) experimental building (Fig. 1(a)), located in the suburb of Lausanne, Switzerland (46°31′17″N, 6°34′02″E, alt. 396 m.). In every office, occupants have the possibility to tilt or open up to any angle each of the two

[1] An additional effect from season was noticed (e.g. similar outdoor temperatures do not imply same action probabilities in spring or autumn), although this is not considered in the final model.

General view of the south façade of the LESO building

Typical cellular office of the LESO building

Fig. 1. Features of the LESO building.

windows (height 90 cm, width 70 cm). Furthermore, external lower and upper roller blinds are controllable within each office. Six offices are occupied by two persons, which can both individually access their own window, while eight offices accommodate single occupant able to act on the two windows. It is safe to leave windows open (eg. for night ventilation) during periods of absence, except on the ground floor. A typical office is shown in Fig. 1(b).

All 14 south-facing cellular offices of this building have been equipped with sensors whose real-time measurements are archived by a centralised EIB data acquisition system. For a period covering 19 December 2001–15 November 2008 (with the exception of a few short interruptions caused by maintenance and technical reasons), the following variables were continuously[2] measured:

- Local indoor temperature (noted θ_{in} hereafter), by Pt-100 resistance thermometers;
- Occupancy, by infrared detectors;
- Window openings and closings by microswitches on each of the two windows.[3]

Outdoor temperature (noted θ_{out} hereafter) has been measured since 17 March 2005 by a sensor located on the roof. In parallel, a weather station located 7.7 km away has recorded measurements every 10 min of temperature, mean wind speed (v_{wind}) and direction (α_{wind}), relative humidity (φ_{out}) and rainfall (D_{prec}). A statistical summary of all the above variables is presented in Table 1, where entries for occupancy and window openings refer to real-time proportions of occupied offices and open windows.

As noted above, local outdoor climate data are missing for the first three years of measurements. To rectify this, linear regression between local and meteorological data for the period with data was used to extrapolate from meteorological measurements for the period without local data, with a good agreement ($R^2 = 0.96$).

The measurements of wind speed and direction present the additional problem of the highly local nature of observations in built-up settings, which undermines the relevance of more distant observations. We may however reasonably use a coarse representation of wind speed and direction, by considering four levels of wind intensity defined by the observed quartiles of wind speed in the weather station given in Table 1. We similarly use four levels for wind direction, defined by 90° domains centered along the cardinal points. These choices allow us to assess the existence of an effect of wind on window opening behaviour (but not to quantitatively estimate its influence).

3. Modelling occupants' adaptive actions

We present here three mathematical methods for the modelling of occupants' adaptive actions. These include logistic regression, discrete-time Markov processes and survival analysis applied to continuous-time random processes.

3.1. Theory of logistic models

When modelling human actions, it is of interest to predict whether an action has taken place, given a set of independent variables. Formally, this implies the inferring of a relationship between a dichotomous outcome variable Y and a set of p independent predictors (covariates) $\mathbf{x} = (x_1,\dots,x_p)$. In our case we will set $Y = 1$ if a window is open and $\mathbf{x}$ may potentially include all driving variables available. Classical least squares regression theory used for linear models is inappropriate for binary outcome variables, because of the violation of the crucial assumption that errors are normally distributed. Indeed for binary data, it is straightforward to show that the residuals $\varepsilon = y - y_{fitted}$ are distributed according to the binomial distribution.

In order to overcome the above limitation, the class of *generalised linear models*[4] (GLM) was developed. In our case of a binary outcome variable Y, the analysis is performed using the binomial family of GLMs. The quantity of interest is the mean value of the outcome variable (in other words the probability that the outcome binary variable Y will be one), given the values of a set of independent variables $\mathbf{x} = (x_1,\dots,x_p)$. This quantity is called the *conditional mean* $E(Y|\mathbf{x})$ and we will set $p(\mathbf{x}) = E(Y|\mathbf{x})$ to simplify the notation. In classical linear models we would assume that $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, but it would then be possible for p to take values outside the interval [0, 1]. It is thus necessary to use a suitable transformation g of $p(\mathbf{x})$. A classical choice is the canonical logit transformation $g(p) = \log(p(\mathbf{x})/(1 - p(\mathbf{x})))$, which defines the class of *logistic regression models*.[5] In this case the probability distribution is called the *logit distribution*, defined as:

[2] The acquisition system records the changes observed by measurement devices in real time. Temperature is recorded along with a time stamp once the variation exceeds 0.06 °C.

[3] These devices do not record the opening angle and do not distinguish tilted openings.

[4] A detailed theoretical background on generalised linear models is available in [28].

[5] More specialised developments on logistic regression may be found in [29].

Table 1
Descriptive statistics of measured parameters.

Variable	Min.	$q(25\%)$	Median	Mean	$q(75\%)$	Max.
Occupancy	0%	0%	0%	16.0%	28.6%	100%
Openings	0%	0%	7.1%	14.9%	0.25%	71.4%
θ_{in} [°C]	13.8	22.2	23.3	23.35	24.5	31.1
θ_{out} [°C]	−9.7	5.7	12.1	12.24	18.4	37.1
φ_{out} [%]	10.5	57.0	68.6	67.41	79.3	100.0
v_{wind} [m/s]	0.0	1.5	2.5	3.304	4.7	16.8
α_{wind} [°]	0	46	168	155.8	236	360
D_{prec} [mm/h]	0	0	0	0.1115	0	96.6

$$p(x_1,\dots,x_p) \;=\; \frac{\exp(\beta_0+\beta_1x_1+\ldots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\ldots+\beta_px_p)}, \tag{1}$$

where $\beta = (\beta_0,\dots,\beta_p)$ are constants estimated by regression through maximum likelihood estimation. We will often refer to the parameter β_0 as the *intercept* and β_i ($i \neq 0$) as the *slope* associated with the variable x_i.

We present the statistical tests used for logistic regression in Appendix A. It is nevertheless useful to note here some of the properties of the logit distribution with a single variable x, intercept β_0 and slope β_1. In the case $\beta_1 > 0$, we notice that $p(x)$ is monotonously increasing, $p(x) \to 0$ for small x, and $p(x) \to 1$ for large x (and conversely if $\beta_1 < 0$). There is a *characteristic value* $x_{50} = -\beta_0/\beta_1$ for which $p(x) = 0.5$, and the variation of p is proportional to β_1 at this point, as $\mathrm{d}p/\mathrm{d}x(x_{50}) = \beta_1/4$.

A limitation of using a single probability distribution is that it does not predict a formal time-evolving probability for an action to be performed, based on a set of given environmental conditions. Its purpose is to predict, based on the included variables, the probability for the outcome variable Y to take the value one (or the system of interest to be found in this state), rather than for the transition of this variable between states. Such a distribution thus does not explicitly provide any probability of direct action and therefore does not describe the real dynamic processes of the system to be modelled.

3.2. Theory of discrete-time Markov processes

In order to overcome the above limitation, a possibility is to infer a stochastic process providing transition probabilities between the states of a window, modelled as a random variable Y_t, for a suitably chosen time step. A possibility would be to restrict ourselves to Markov processes, which consider only the previous state and are thus independent of the process history (Markov property). In this case, we have that:

$$P(Y_{t+1}=j|Y_t=i,Y_{t-1}=h,\ldots)(\mathbf{x}) = P(Y_{t+1}=j|Y_t=i)\,(\mathbf{x}) = P_{ij}(\mathbf{x}). \tag{2}$$

We call $P_{ij}(\mathbf{x})$ the *probability of transition* from state i to state j, as a function of parameters $\mathbf{x} = (x_1,\dots,x_p)$.

One of the main drawbacks of such discrete-time random processes is the necessity to choose a fixed time step, leading to a possible loss of information (e.g. short duration openings could be ignored if they last less than the given time step), or to redundant calculations (e.g. for periods in which there is no change of state). Furthermore the derived probability transitions remain time step dependent, which limits somewhat their generality. Thus, the choice of the time step must be carefully determined.

In our case, Y_t is defined as the time-evolving state of a window and $P_{ij}(\mathbf{x})$ is a 2×2 matrix, as we consider two possible states for the window (open and closed).

3.3. Theory of survival analysis

This approach has not yet been used in our context, but has long-since been applied in reliability studies and biomedical research. Survival analysis typically attempts to model distributions of survival durations until a "failure" occurs.[6] Thus, this approach could be used to model the duration for which an occupant will leave a window closed following their arrival, or open since its opening, according to relevant driving variables. The durations t are modelled as a non-negative random variable T, with *distribution function* $F(t) = P(T \le t)$ and corresponding *probability density function* $f(t)$. We define then the associated *survival function* (or *reliability function*) $S(t) = P(T \ge t) = 1 - F(t)$ and the *hazard function* $h(t) = f(t)/S(t)$. It follows from these definitions that $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$ and $h(t) = -\mathrm{d}\log(S(t))/\mathrm{d}t$.

A strength of these statistical methods lies in the possibility of including survival times containing partial information such as survival for at least a given period of time, without needing further knowledge about the process after this time. The inclusion of these so called *censored data* allows for non-comprehensive observations, for instance such as closing on departure or the absence of opening during the whole occupancy period, to be included into the modelling of duration of window openings. Provided r_i the number of surviving elements until time t_i (including censored observations), and d_i the number of "failures" at time t_i, we find that the conditional probability to survive beyond t_i, knowing that the subject is alive just before t_i is given by $(r_i - d_i)/r_i$. The survival function can then be inferred by the non-parametric *Kaplan–Meier estimator*,

$$\widehat{S}_{KM}(t) \;=\; \prod_{i|t_i<t}\left(1-\frac{d_i}{r_i}\right), \tag{3}$$

which defines unambiguously an estimator of the distribution $F(t) = 1 - S(t)$.

However, in order to perform a parametric estimation, a survival distribution has to be fitted using maximum likelihood regression. The reader may consult [30] for further details. We consider here two particular distributions to be used in our study:

- The *exponential distribution*, with $f(t) = \lambda\exp(-\lambda t)$, $S(t) = \exp(-\lambda t)$, which assumes a constant hazard rate $h(t) = \lambda = \text{const}$, is the simplest model.
- The *Weibull distribution*, with $f(t) = \lambda\alpha(\lambda t)^{\alpha-1}\exp(-(\lambda t)^{\alpha})$, $S(t) = \exp(-(\lambda t)^{\alpha})$ and $h(t) = \lambda\alpha(\lambda t)^{\alpha-1}$, where the parameter α is called the *shape* and λ the *scale*, offers more flexibility. In the special case $\alpha = 1$, it corresponds to the exponential distribution. When $\alpha < 1$, $h(t)$ decreases with t, and failures are more likely at small times, and conversely for $\alpha > 1$.

It is worth noting that a random variable T has an exponential distribution $S(t) = \exp(-\lambda t)$ if and only if it has the memoryless property $P(T > s + t|T > s) = P(T > t)$ for all s, $t \ge 0$. In other words, a Markov process may be rigorously used if and only if delays between actions can be correctly modelled by an exponential distribution.

[6] For example, typical applications include the study of survival times of leukemia patients until death (with assessment of the effect of differentiated treatments); or the prediction of the mean lifetime of industrial components in different conditions.

4. Results

We present in this section some preliminary observations on window opening behaviour followed by results obtained from application of the different models inferred on the basis of the concepts introduced in Section 3. The statistical software R [31] was used for all data analyses and for programming the models.

4.1. Preliminary observations

Occupancy patterns: As noticed in previous surveys [4,22,25], actions on windows most commonly occur when occupants arrive or leave their offices (Fig. 2). For this a first difficulty is the arbitrary choice of a temporal threshold for the definition of events occurring just after arrival or before departure. We plot the empirical cumulative distribution functions (ECDF) of time intervals between occupancy transitions and actions on windows (Fig. 2(a)), and we notice that a threshold of 5 min defines a limit above which the slopes of the ECDFs remain relatively constant, suggesting that all events related to occupants' arrival and departure have by that time been captured. A possible explanation is that during this time occupants may perceive considerable differences in thermal and/or olfactory stimuli compared to their previous (possibly external) environment and their offices. These differences may be exacerbated by the lack of adaptive actions while the occupant was not present.

For offices with two occupants, we notice a slightly greater proportion of openings during occupancy (Fig. 2(b)). However, in this case many actions on arrival or departure are classified as intermediate actions, for instance when an occupant arrives in an already occupied office and opens the window. This question relates to the issue of group actions, discussed in Section 4.5.

Influence of environmental parameters: Our first step in the inference of a model linking the occurrence of open windows and one or several physical predictors involves examining the observed proportion of windows open with respect to the measured physical parameters. In Fig. 3 we chart separately these proportions (based on observations taken every 10 min) grouped by bins of each measured physical parameter.

A clearly increasing proportion may be observed for θ_{in} rising from 20 °C to 28 °C, with a possibly significant decrease above this range[7] (Fig. 3(a)). A similar phenomenon occurs with θ_{out}, the maximum proportion being reached around 26 °C, above which a decrease is clearly significant (Fig. 3(b)). This type of behaviour was previously remarked by Rijal et al. [23]. Both thermal variables are thus clearly linked with actions on windows, in agreement with previous research. A less sharp decrease in the proportion of windows open is visible when outdoor humidity rises (Fig. 3(c)). Further examination will assess whether these variations are intrinsically influenced by each parameter, as these variables are correlated ($\rho(\theta_{in},\theta_{out}) = 0.62$, $\rho(\theta_{out},\varphi_{out}) = -0.45$).

Mean wind speed v_{wind} (Fig. 3(e)) increase is linked with a decreased proportion of open windows for $v_{wind} > 2$ m/s. No particular variation may be observed with wind direction (Fig. 3(f)). No clear pattern may be noticed with respect to rainfall D_{prec} (Fig. 3(d)), which is correlated with relative humidity: $\rho(D_{prec},\varphi_{out}) = 0.25$.

Based on these preliminary examinations, a relevant model would include in order of priority θ_{in}, θ_{out} and possibly φ_{out} and v_{wind}.

Variability between occupants and personal patterns related to actions on windows: We have noticed that, while the climatic conditions are fairly similar between the studied offices, some occupants use their windows more frequently than do others. Reinhart [26] and Rijal et al. [16] suggested a distinction between "active" and "passive" occupants in the case of actions on blinds and lights; perhaps a similar classification may be applied for windows? To test this hypothesis we shall first produce generalist models based on all occupants, and go on to develop a method to treat variability among occupants in Section 4.5.

Towards an appropriate model for the prediction of actions on windows: These initial observations enable us to draw some first conclusions about the general form of a comprehensive model:

- Occupancy patterns should be integrated, for instance to facilitate separate treatment of actions on arrival, departure and during occupancy. This implies the necessity to explicitly model occupancy itself;
- Among possible other parameters, indoor and outdoor temperatures are the main driving stimuli for actions on windows;
- A possible refinement could be to distinguish between "active" and "passive" users, and possibly between single and multi-occupied offices.

4.2. Models based on logit distributions

From now on we use the following notation for all the models based on (linear) logit distributions:

$$\begin{aligned}\text{logit}(p) &= \log\left(\frac{p}{1-p}\right) = a + b_{in}\theta_{in} + b_{out}\theta_{out} + b_{\phi}\phi_{out} \\ &\quad + b_R f_R + b_{WS} f_{WS} + b_{WD} f_{WD},\end{aligned} \tag{4}$$

where a and b_i are the regression parameters and f_R, f_{WS} and f_{WD} are dummy variables (based on a crude discretisation) for respectively rain presence, wind speed levels and wind direction sectors.

4.2.1. Univariate logistic models

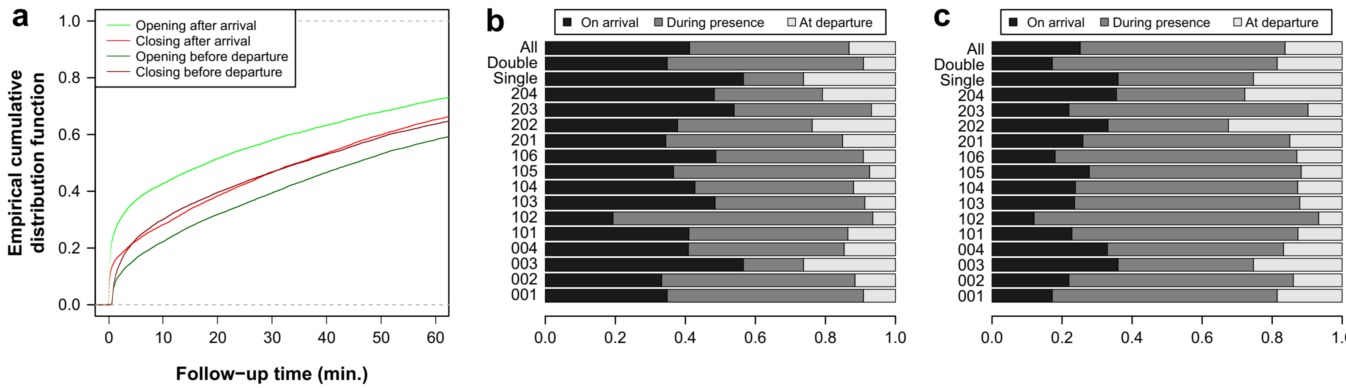
In this section, we present results from separate logistic regressions using each available independent variable, together with some possible transformations of these latter.

Models with untransformed variables: The regression curves are presented in Fig. 3(a)–(f) and the regression parameters are given in Table 2. From this we observe statistical significance ($p < 0.001$) of each of the variables tested. The model with θ_{out} has the largest likelihood ratio statistic, implying that it best describes the variations of our outcome variable.

However, as discussed in Appendix A, statistical significance itself does not necessarily provide clear-cut conclusions concerning the model's capacity to correctly explain our outcome variable. We therefore give in Table 3 a summary of the possible criteria of goodness-of-fit for each of these models. According to these goodness-of-fit criteria, the model with θ_{out} once again offers the best fit among all variables. We thus conclude that θ_{out} should be integrated in a final model, possibly in conjunction with other variables if their contributions are statistically significant and improve the quality of adjustment. The implications of this superiority of θ_{out} as a predictive variable are discussed in Section 5.1.

Models using polynomial logits: We noticed in Fig. 3(b) that a linear logit does not predict well the observed proportions of windows open at high outdoor temperatures. In order to account for this phenomenon, a possible refinement would be to use a polynomial of degree q for the logit of the probability. In this case:

[7] A higher proportion of windows open at low θ_{in} is noticeable in Fig. 3(a). This is caused by a window left open while θ_{out} is low. The 95% confidence intervals are large, owing to the very small number of observations within the relevant bins.



a: ECDF of time intervals between occupancy transitions and actions on windows

b: Occupancy-dependent repartition of window openings for each office and for all single (S) and all double (D) occupied offices

c: Occupancy-dependent repartition of window closings for each office and for all single (S) and all double (D) occupied offices

Fig. 2. Observed proportions of window openings and closings taking place in different occupancy situations (dark grey: arrival, grey: during occupancy, light grey: departure).

$$\mathrm{logit}(p) = a + b_1\theta_{\mathrm{out}} + b_2\theta_{\mathrm{out}}^2 + \cdots + b_q\theta_{\mathrm{out}}^q, \tag{5}$$

where we use stepwise logistic regression to determine the highest significant order. This procedure determines that a fourth degree polynomial is appropriate, with regression parameters $a = -2.275 \pm 0.008$, $b_1 = (5.45 \pm 0.23) \times 10^{-2}$, $b_2 = (-0.70 \pm 3.24) \times 10^{-4}$, $b_3 = (3.86 \pm 0.17) \times 10^{-4}$, $b_4 = (-1.112 \pm 0.029) \times 10^{-5}$. We see in Fig. 3(g) that the associated probability distribution fits better the observed proportions; particularly for high values of θ_{out}. Furthermore, regressions with polynomials of lower degree do not offer clear improvements compared to the linear logit model. Although goodness-of-fit indicators are not much improved (Table 3), this is the best model when using one sole predictor.

Models based on deviations from comfort temperature: Another possible choice for a driving variable is to use the deviation between θ_{in} or θ_{out} and a comfort temperature $\theta_{\mathrm{in,comf}}$; for example defined by the adaptive comfort model of de Dear and Brager [32]. We set thus $\theta_{\mathrm{in,comf}}^{(\mathrm{DB})} = 17.8 + 0.31\theta_{\mathrm{out,mm}}$, where $\theta_{\mathrm{out,mm}}$ is the monthly mean of outdoor temperature. We perform logistic regression with $(\theta - \theta_{\mathrm{in,comf}}^{(\mathrm{DB})})$ as a driving variable, alternatively with $\theta = \theta_{\mathrm{in}}$ and $\theta = \theta_{\mathrm{out}}$. The corresponding results are given in Fig. 3(h) and Table 3 (bottom).

In this case, we obtain slightly lower goodness-of-fit and likelihood ratio; that is, the quality of adjustment is somewhat lower than when using raw thermal variables. It is however worth noting that the proportion of windows open reaches a maximum near $\theta_{\mathrm{out}} = (\theta - \theta_{\mathrm{in,comf}}^{(\mathrm{DB})})$. This procedure was repeated with the equation given by the CEN standard (see Section 1), producing similar results.

4.2.2. Multivariate logistic models

Following from these univariate models, we proceed to consider models with several variables and assess their increased predictive value. In this we determine the best model containing two variables, and provided the significance of the added variable and the stability of the primary variable; continuing this procedure for other predictors until no added significance is obtained. This procedure is known as *forward selection.*

Models with two variables: Based on logistic regression for models including together θ_{out} and each other available variable, we observe (Table 3) that the model with θ_{out} and θ_{in} ($a = 1.459 \pm 0.032$, $b_{\mathrm{out}} = 0.14477 \pm 0.00033$, $b_{\mathrm{in}} = -0.1814 \pm 0.0015$) has the highest statistical significance, according to the likelihood ratio statistic. Furthermore, this model is the one that fits best the data, according to all our statistical criteria; but the improvement to these indicators from adding θ_{in} is rather modest. However a plot of the observed proportions of windows open versus θ_{out} and θ_{in}, with regression surface levels (Fig. 3(i)), shows that observed variations are better accounted for and thus confirms the existence of an independent contribution of each variable. Finally, the stability of the slope associated with θ_{out} is preserved, as its standard error remains extremely low, which shows that the correlation between θ_{in} and θ_{out} is not problematic for this model.

Models with three or more variables: Now that the model including θ_{out} and θ_{in} is retained, we check for the significance of the inclusion of a third parameter. Based on regression results for the models with a third variable, the best model includes the external relative humidity φ_{out} and this inclusion is statistically significant ($p < 0.001$). However the goodness-of-fit criteria increase only very slightly (Table 3); that is, the added predictive accuracy from the inclusion of φ_{out} is marginal. Some other parameters in models with four or five variables were also found to be statistically significant, but without any increase in the goodness-of-fit indicators. For the sake of parsimony, it seems sensible to keep the model with just the two variables θ_{out} and θ_{in}.

Other stimuli: Inspired by the results of Herkel et al. [25], we attempted to include a factor with twelve levels corresponding to each month of the year, in order to check the existence of an additional effect of season on window actions. This was not found to bring any significant improvement; that is we observe almost the same logit distributions based on θ_{out} for every month.

4.3. Model based on a discrete-time Markov process

As noted earlier, a single probability distribution ignores the real dynamic processes leading occupants to perform actions, as the data used to infer them are aggregated observations of window states, but not actual opening or closing actions.[8] In other words these models do not describe an actual probability of opening or closing, but a probability for a window to be "found" open, provided relevant physical parameters. Furthermore it ignores the particular patterns caused by occupancy events, like arrivals or departures of occupants. We thus present in this section an

[8] The approach used in the Humphreys algorithm [16,23] is a possible adjustment choice to include dynamics in such probability distributions, although not based on observed actions.

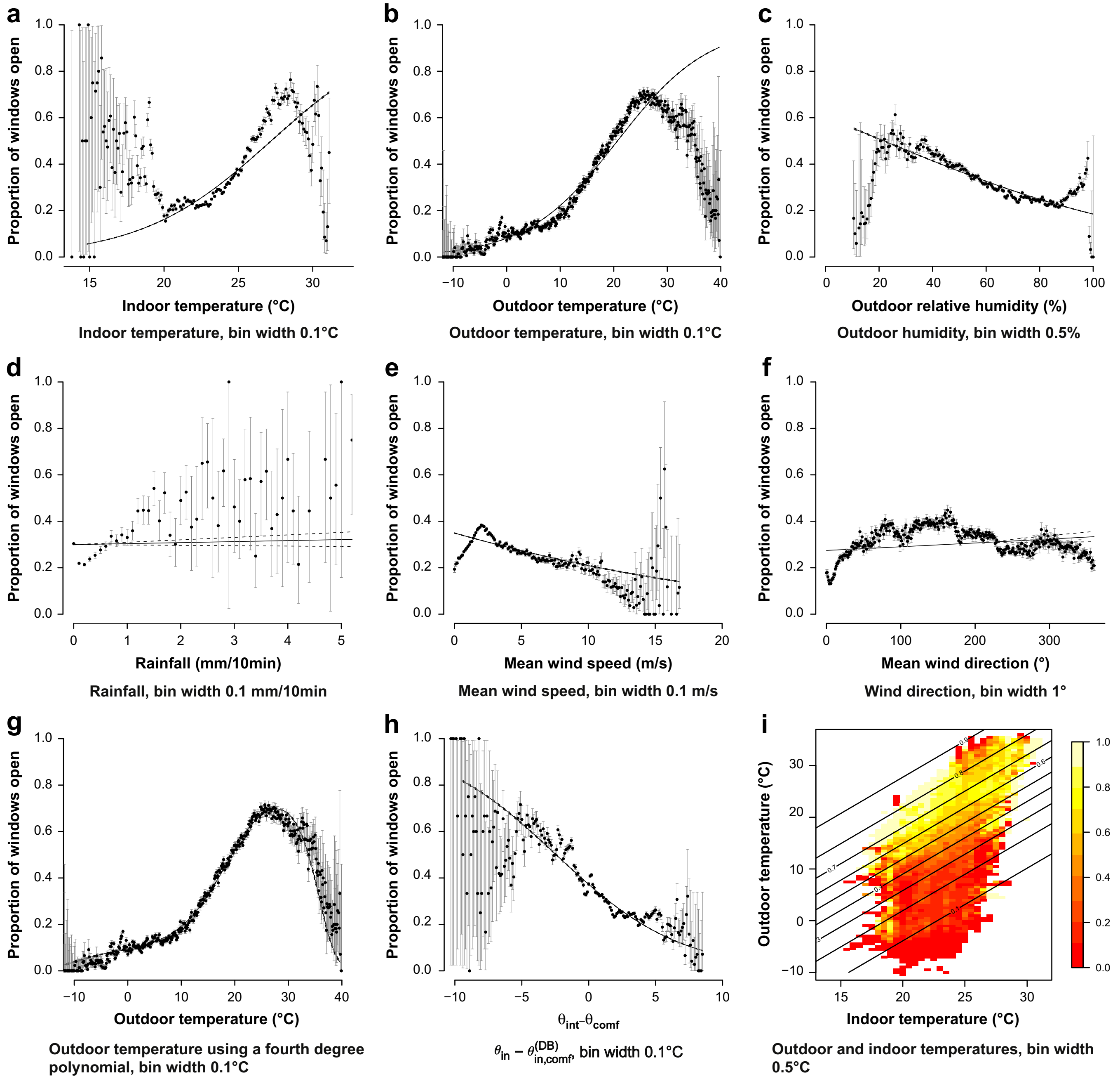


Fig. 3. Observed proportion of windows open for specified bandwidths as a function of different measured physical parameters: (a)–(h) with binomial 95% level confidence intervals and logistic regression curves, (i) with contour lines of equal fitted probabilities curves.

alternative dynamic modelling approach to account for the real adaptive processes of occupants.

Guided by the initial observation that occupancy events have an influence on actions (Section 4.1), we may infer different transition probabilities P_{ij} for these events, so that we have three different sub-models for actions on arrival, at departure and during occupancy, as proposed in [22,25]. Simulation may then be conducted as presented in Fig. 4: opening on arrival is predicted by a specific probability $P_{01,\mathrm{arr}}$, and closing on arrival by $P_{10,\mathrm{arr}}$. Actions after arrival are predicted by another sub-model launched at regular time steps, with transition probabilities $P_{01,\mathrm{int}}$ if the window is closed at this time and $P_{10,\mathrm{int}}$ if it is opened. When the occupant leaves his office, a third sub-model predicts actions on departure, with transition probabilities $P_{01,\mathrm{dep}}$ and $P_{10,\mathrm{dep}}$. In each case, P_{00} and P_{11} are easily deduced: $P_{00} = 1 - P_{01}$ and $P_{11} = 1 - P_{10}$.

For each sub-model, we filter the data to retain observations related to the relevant occupancy status, and perform logistic regressions on the most relevant environmental parameters; retaining the optimal set of variables in this adapted version of Equation (4):

$$\begin{aligned}\operatorname{logit}(p) = {} & a + b_{\mathrm{in}}\theta_{\mathrm{in}} + b_{\mathrm{out}}\theta_{\mathrm{out}} + b_{\mathrm{dm}}\theta_{\mathrm{out,dm}} + b_{\mathrm{R}}f_{\mathrm{R}} + b_{\mathrm{WS}}f_{\mathrm{WS}} \\ & + b_{\mathrm{WD}}f_{\mathrm{WD}} + b_{\mathrm{GF}}f_{\mathrm{GF}} + b_{\mathrm{pres}}T_{\mathrm{pres}} + b_{\mathrm{abs,prev}}f_{\mathrm{abs,prev}} \\ & + b_{\mathrm{abs,next}}f_{\mathrm{abs,next}},\end{aligned} \tag{6}$$

Table 2
Regression parameters for logistic models including a single variable; in the case of models including factors with multiple levels, the intercept and likelihood ratio are provided in the row of the chosen reference level.

Variable	a	Z	b	Z	L.R.
θ_{out}	-2.472 ± 0.0045	-5532.7	0.1212 ± 0.00027	454.4	264262
θ_{in}	-6.22 ± 0.026	-237.2	0.230 ± 0.0011	207.3	45138
ϕ_{out}	0.381 ± 0.0083	46.1	-0.0180 ± 0.00012	-148.0	22119
v_{wind}	-0.582 ± 0.0031	-185.0	-0.0754 ± 0.00081	-92.7	8936
α_{wind}	-0.990 ± 0.0034	-290.0	0.00104 ± 0.000018	58.4	3421
D_{prec}	-0.828 ± 0.0019	-442.7	0.15 ± 0.014	11.0	122
f_R	-0.809 ± 0.0019	-424.1	-0.255 ± 0.0079	-32.3	1079
f_{WS} (<1.5)	-0.883 ± 0.0039	-228.7			15743
f_{WS} (1.5–2.5)			0.368 ± 0.0052	70.7	
f_{WS} (2.5–4.7)			0.102 ± 0.0053	19.2	
f_{WS} (>4.7)			-0.279 ± 0.0055	-50.6	
f_{WD} (North)	-1.214 ± 0.0036	-340.3			22782
f_{WD} (East)			0.630 ± 0.0056	112.6	
f_{WD} (South)			0.673 ± 0.0049	137.2	
f_{WD} (West)			0.348 ± 0.0053	66.1	
$\theta_{out} - \theta_{comf}^{(DB)}$	0.266 ± 0.0031	85.8	0.1310 ± 0.00031	424.0	226071
$\theta_{in} - \theta_{comf}^{(DB)}$	-0.522 ± 0.0025	-210.4	-0.215 ± 0.0011	-196.1	39808
$\theta_{out} - \theta_{comf}^{(CEN)}$	0.475 ± 0.0035	137.4	0.1320 ± 0.00031	424.7	228533
$\theta_{in} - \theta_{comf}^{(CEN)}$	-0.861 ± 0.0020	-440.9	-0.251 ± 0.0012	-203.04	42952

where T_{pres} is the ongoing presence duration, $f_{abs,prev}$, $f_{abs,next}$ and f_{GF} are binary variables equal to one respectively for preceding or following absences longer than 8 h and for offices not on ground floors, and b_{pres}, $b_{abs,prev}$, $b_{abs,next}$, b_{GF} are their associated regression parameters.

4.3.1. Sub-model for actions on arrival

Based on our preliminary observations, we will include all actions performed within 5 min of arrival in this sub-model. We notice that 2.3% of the arrivals with windows closed were followed by an opening, while 3.1% of the arrivals with windows open resulted in a closing. Based on the variable selection procedure presented in Section 4.2.2, we conclude that the best model with a single variable uses θ_{in} as a predictor for openings and for closings, while the second most influential variable is θ_{out}. We show regression results in Table 4 (top) and in Fig. 5(a)–(d) the observed proportions of actions with contour levels of regression surface. From these results it is clear that θ_{in} exerts the dominant influence on both the opening and closing of windows on arrival.

Table 3
Goodness-of-fit estimators (see Appendix A) for logistic models including one or several variables; the model with θ_{out} and ϕ_{out} did not converge.

Variables	AUC	R^2	B	D_{xy}
θ_{out}	0.769	0.247	0.172	0.537
θ_{in}	0.611	0.046	0.204	0.222
φ_{out}	0.577	0.022	0.208	0.154
v_{wind}	0.536	0.009	0.211	0.072
α_{wind}	0.531	0.004	0.211	0.061
D_{prec}	0.493	0.000	0.212	0.000
f_R	0.507	0.001	0.212	0.015
f_{WS}	0.564	0.016	0.209	0.128
f_{WD}	0.575	0.023	0.208	0.151
θ_{out} (polyn.)	0.770	0.262	0.169	0.540
$\theta_{out} - \theta_{comf}^{(DB)}$	0.758	0.226	0.174	0.516
$\theta_{in} - \theta_{comf}^{(DB)}$	0.604	0.043	0.203	0.208
$\theta_{out} - \theta_{comf}^{(CEN)}$	0.759	0.228	0.173	0.518
$\theta_{in} - \theta_{comf}^{(CEN)}$	0.607	0.046	0.203	0.214
θ_{out}, θ_{in}	0.774	0.260	0.170	0.547
θ_{out}, v_{wind}	0.769	0.247	0.171	0.538
θ_{out}, α_{wind}	0.770	0.249	0.171	0.540
θ_{out}, D_{prec}	0.769	0.247	0.171	0.538
θ_{out}, f_R	0.769	0.247	0.172	0.538
θ_{out}, f_{WS}	0.769	0.247	0.171	0.538
θ_{out}, f_{WD}	0.770	0.248	0.171	0.539
θ_{out}, θ_{in}, φ_{out}	0.777	0.268	0.168	0.554

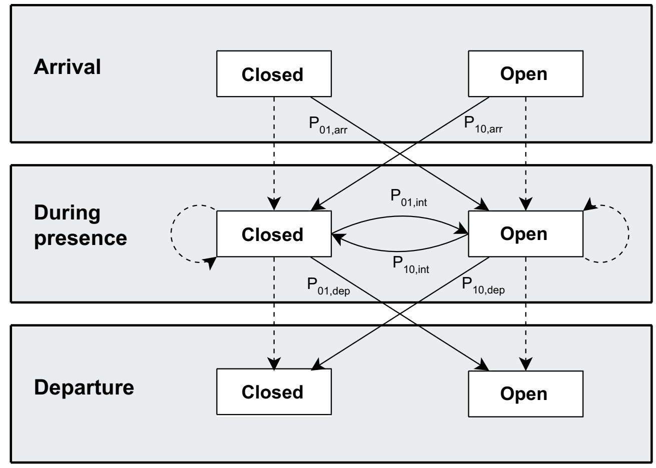


Fig. 4. General scheme of the Markov process.

For $P_{01,arr}$, the presence of rain is a significant factor, but prior absence duration has a stronger influence. With these variables, Equation (6) becomes:

$$\operatorname{logit}\left(P_{01,arr}\left(\theta_{in}, \theta_{out}, f_{abs,prev}, f_R\right)\right) = a + b_{in}\theta_{in} + b_{out}\theta_{out} + b_{abs,prev}f_{abs,prev} + b_R f_R. \quad (7)$$

On the contrary, we have not found any other significant variable for $P_{10,arr}$. The final model is thus:

$$\operatorname{logit}(P_{10,arr}(\theta_{in}, \theta_{out})) = a + b_{in}\theta_{in} + b_{out}\theta_{out}. \quad (8)$$

Goodness-of-fit indicators are provided in Table 5. Once again there are noticeable differences in behaviour between offices, which will be discussed in Section 4.5.

4.3.2. Sub-model for actions during occupancy

Among the 5-min time steps (excluding arrival and departure periods) starting with closed windows, 0.4% end up with an opening, while 1.1% of those beginning with open windows include a closing event. Actions during occupancy are thus extremely rare; particularly openings.

We see that θ_{in} is the main driving variable for $P_{01,int}$, while θ_{out} dominates for $P_{10,int}$ (indeed θ_{in} is barely significant in this case). This shows that θ_{in} is the real underlying stimulus for openings, while θ_{out} (linked to the feedback of the opening) determines primarily the probability of closing (e.g. to prevent over- or underheating), see Fig. 5(b)–(e).

Rainfall and wind are not significant for openings during occupancy. Wind speed alone is not significant for closings, but for south and east wind directions the probability of closing our south-facing window is increased and reduced for the westerly direction (taking north as reference level[9]). However this refinement does not result in improved adjustment quality, so once again for parsimony we do not include this variable. Both current occupancy duration and the

[9] Setting $b_{WD,north} = 0$, we get $b_{WD,east} = 0.235 \pm 0.048$, $b_{WD,south} = 0.212 \pm 0.044$, $b_{WD,west} = -0.113 \pm 0.046$.

Table 4
Regression parameters for final Markovian transition probabilities.

Type	Parameters	Opening probabilities			Closing probabilities		
		Estimate	Z	χ^2	Estimate	Z	χ^2
Arrival	a	-13.70 ± 0.40	-27.84		3.95 ± 0.39	10.11	
	b_{in}	0.308 ± 0.017	17.67	312.21	-0.286 ± 0.018	-15.72	247.24
	b_{out}	0.0395 ± 0.0036	10.86	117.95	-0.0500 ± 0.0049	-10.27	105.55
	$b_{abs,\ prev}$	1.826 ± 0.048	37.80	1428.6			
	b_R	-0.43 ± 0.12	-3.59	12.86			
Interm.	a	-11.78 ± 0.30	-38.83		-4.14 ± 0.24	-17.57	
	b_{in}	0.263 ± 0.014	19.44	377.75	0.026 ± 0.011	2.34	5.48
	b_{out}	0.0394 ± 0.0036	10.86	95.01	-0.0625 ± 0.0024	-25.85	668.27
	b_{pres}	$(-9.00 \pm 0.57) \times 10^{-4}$	-15.68	245.81			
	b_R	-0.336 ± 0.088	-3.83	14.65			
Departure	a	-8.72 ± 0.23	-37.19		-8.68 ± 0.25	-16.61	
	b_{in}				0.222 ± 0.024	9.28	86.04
	$b_{out,dm}$	0.1352 ± 0.0078	17.36	301.31	-0.0936 ± 0.0067	-14.06	197.60
	$b_{abs,next}$	0.85 ± 0.12	5.79	46.51	1.534 ± 0.077	19.99	399.79
	b_{GF}	0.82 ± 0.14	6.82	33.47	-0.845 ± 0.074	-11.48	131.79

occurrence of rain are found to be significant for $P_{01,int}$. We thus have the following models,

$$\mathrm{logit}\left(P_{01,int}\left(\theta_{in},\theta_{out},f_{abs,prev},f_R\right)\right) = a + b_{in}\theta_{in} + b_{out}\theta_{out} + b_{pres}T_{pres} + b_R f_R, \tag{9}$$

$$\mathrm{logit}(P_{10,int}(\theta_{in},\theta_{out})) = a + b_{in}\theta_{in} + b_{out}\theta_{out}, \tag{10}$$

with regression parameters given in Table 4 (middle). Goodness-of-fit indicators are lower than for the sub-model relating to actions on arrival (Table 5). One final observation is that transition probabilities remain in both cases very close to zero, and thus consecutive repeated predictions of the same state are very likely so that the associated computation is wasteful. An alternative would be to increase the time step, but this would result in neglecting openings of short duration or artificially increasing the duration of other openings. A more appropriate method for intermediate actions is proposed in Section 4.4.

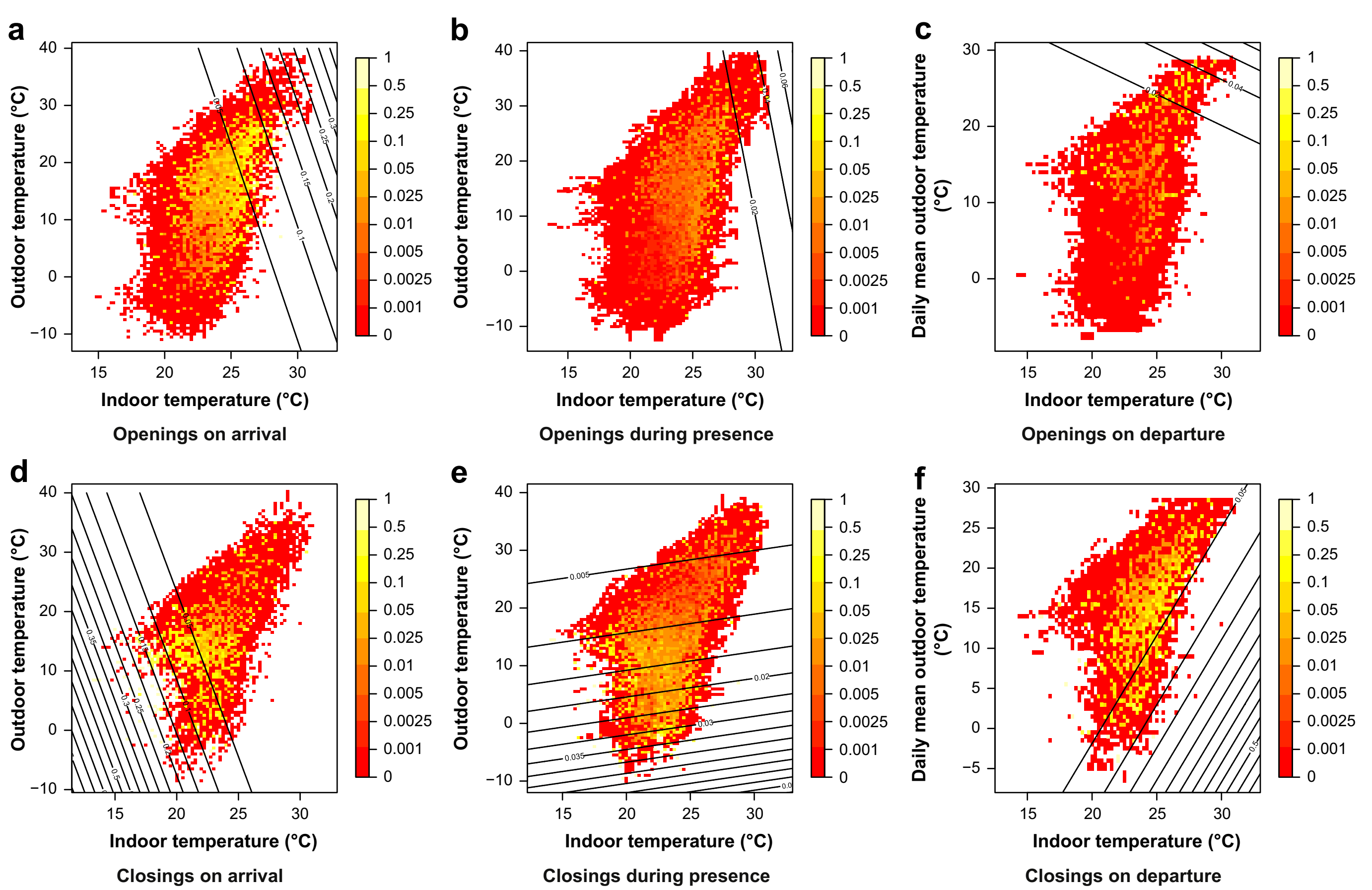


Fig. 5. Observed transition probabilites (given on a quasi-logarithmic scale), versus bins of indoor and outdoor temperature, with contour lines of equal fitted probabilities.

Table 5
Goodness-of-fit estimators for Markovian transition probabilities.

Model, variables	AUC	R^2	B	D_{xy}
$P_{01,arr}$				
θ_{in}	0.687	0.049	0.022	0.374
θ_{out}	0.688	0.042	0.022	0.376
θ_{in}, θ_{out}	0.711	0.055	0.022	0.423
θ_{in}, θ_{out}, f_R	0.712	0.056	0.022	0.424
θ_{in}, θ_{out}, f_{abs}, f_R	0.771	0.124	0.022	0.542
$P_{10,arr}$				
θ_{in}	0.683	0.065	0.029	0.367
θ_{out}	0.704	0.047	0.029	0.409
θ_{in}, θ_{out}	0.716	0.080	0.029	0.432
$P_{01,\ int}$				
θ_{in}	0.653	0.025	0.004	0.305
θ_{out}	0.627	0.018	0.004	0.254
θ_{in}, θ_{out}	0.661	0.027	0.004	0.321
θ_{in}, θ_{out}, f_R	0.663	0.028	0.004	0.325
θ_{in}, θ_{out}, T_{pres}, f_R	0.695	0.040	0.004	0.390
$P_{10,int}$				
θ_{in}	0.583	0.007	0.011	0.167
θ_{out}	0.658	0.023	0.011	0.315
θ_{in}, θ_{out}	0.657	0.024	0.011	0.314
$P_{01,dep}$				
$f_{abs,next}$	0.574	0.003	0.004	0.147
$\theta_{out,dm}$	0.732	0.068	0.004	0.465
θ_{in}, $\theta_{out,dm}$	0.743	0.070	0.004	0.486
θ_{in}, $\theta_{out,dm}$, f_{abs}	0.731	0.077	0.004	0.462
θ_{in}, $\theta_{out,\ dm}$, f_{abs}, f_{GF}	0.744	0.085	0.004	0.487
$\theta_{out,dm}$, f_{abs}, f_{GF}	0.738	0.085	0.004	0.477
$P_{10,dep}$				
$\theta_{out,dm}$	0.605	0.012	0.028	0.210
$f_{abs,next}$	0.693	0.030	0.028	0.386
θ_{in}, $\theta_{out,dm}$	0.653	0.028	0.028	0.305
θ_{in}, $\theta_{out,dm}$, f_{abs}	0.716	0.074	0.027	0.432
θ_{in}, $\theta_{out,dm}$, $f_{abs}f_{GF}$	0.730	0.092	0.027	0.460

4.3.3. Sub-model for actions at departure

We notice that occupants opened their initially closed windows in 0.4% of cases when leaving, whilst for 2.8% of departures open windows were closed.

Actions on departure are of a different nature: their goal is not to modify the indoor environment for immediate further occupancy. They may therefore be influenced by other factors, such as the predicted duration of subsequent absence, the desire to induce night ventilation or by security issues. We thus include the observed subsequent absence duration in our model. For both openings and closings, we find that the daily mean outdoor temperature $\theta_{out,dm}$ fits better the data than θ_{out}, and that the position within building (at or above the ground floor) is a significant parameter. In the case of openings on departure, θ_{in} is not a significant parameter. We retain thus the following transition probabilities,

$$\mathrm{logit}\left(P_{01,dep}\left(\theta_{out,dm}, f_{abs,next}, f_{GF}\right)\right) = a + b_{out,dm}\theta_{out,dm} + b_{abs,next}f_{abs,next} + b_{GF}f_{GF}, \tag{11}$$

$$\mathrm{logit}\left(P_{01,dep}\left(\theta_{in}, \theta_{out,dm}, f_{abs,next}, f_{GF}\right)\right) = a + b_{in}\theta_{in} + b_{out,dm}\theta_{out,dm} + b_{abs,next}f_{abs,next} + b_{GF}f_{GF}, \tag{12}$$

with regression parameters given in Table 4 (bottom). Analysis of deviance shows that the duration of absence and the daily mean outdoor temperature are the most significant factors for openings and for closings.

4.3.4. Remarks

By defining factors for previous and next absence durations, our final sub-models for actions on arrival and at departure are in fact divided into separate models for short and long absences. We notice that the start of a long absence period increases the probability of closing at departure, which is an expected result. The fact that a long absence preceding an arrival increases the probability of opening could be explained by the fact that, in this case, odours might have accumulated in the office, which could be partly a consequence that most occupants close their doors when leaving for a long absence. Finally, we have confirmed that thermal stimuli are the key variables influencing actual actions on windows.

Our sub-model for actions on departure takes the following absence duration as input, and thus the future value of a stochastic variable. This implies that a model of occupancy presence should be a pre-process (for the entire simulation period) to the model of window opening.

Goodness-of-fit criteria show that our sub-models do not offer equal performances (Table 5). We obtain the highest predictive power for $P_{01,arr}$, followed by the sub-models for actions on departure and $P_{10,arr}$, and the lowest performance for actions during presence, with $P_{10,int}$ being the least satisfactory sub-model.

4.4. Continuous-time random process

4.4.1. Opening duration

According to the concepts described in Section 3.3, we infer a distribution for the duration during which people leave their window closed following their arrival, and during which the window is left open. Kaplan–Meier estimates of survival curves are shown in Fig. 6(a) and (b), in which each curve refers to an interval of observed initial values of θ_{in} or θ_{out}.

The duration of window openings which were interrupted upon departure needs special treatment. In this case, the reason for closing (or for leaving open) windows is not linked to discomfort, meaning that if the occupant had stayed longer it is not clear that he would have closed the window. We thus have incomplete information: we know that the occupant wished to leave the window open *at least* until this moment. Such opening durations are classified as censored data.

A trend of diminished rates of decay of opening times may be noticed in Fig. 6(a) when θ_{in} rises until 26 °C, while they remain similar above. These decay rates are more clearly differentiated in Fig. 6(b), which implies that opening durations are more strongly associated with θ_{out}. Both variables are significant ($p < 0.001$) according to the log-rank test.

Detailed analysis of the distribution of opening times shows that the hazard rate $h(t)$ is clearly non-constant and decreases with t, meaning that closings have an increased risk of occurring shortly after openings. Based on the 11,870 observed openings and using a Weibull distribution (see Section 3.3) we find that the best model with a single variable uses θ_{out} as its predictor ($p < 0.001$, $R^2 = 0.102$). The estimate for the shape is $\log(1/\alpha) = 0.871 \pm 0.0115$, while the scale is $\lambda = 1/\exp((2.123 \pm 0.071) + (0.1727 \pm 0.0049) \cdot \theta_{out})$. The fitted survival function is shown in Fig. 6(c). The variable θ_{in}, if included with this model, is not statistically significant ($p > 0.1$), likewise other potential variables. These results are consistent with our sub-model for window closings during occupancy developed in Section 4.3, where θ_{out} is the main driving variable in $P_{10,int}$.

4.4.2. Closing duration

The data of closing duration include two types of intervals: delay until opening following occupants' arrival, and delay until opening following a prior closing. We see that θ_{out} has less influence than θ_{in} on closing duration (Fig. 6(d) and 6(e)) and therefore on the decay of survival curves, which differ less in the range of values of θ_{out}

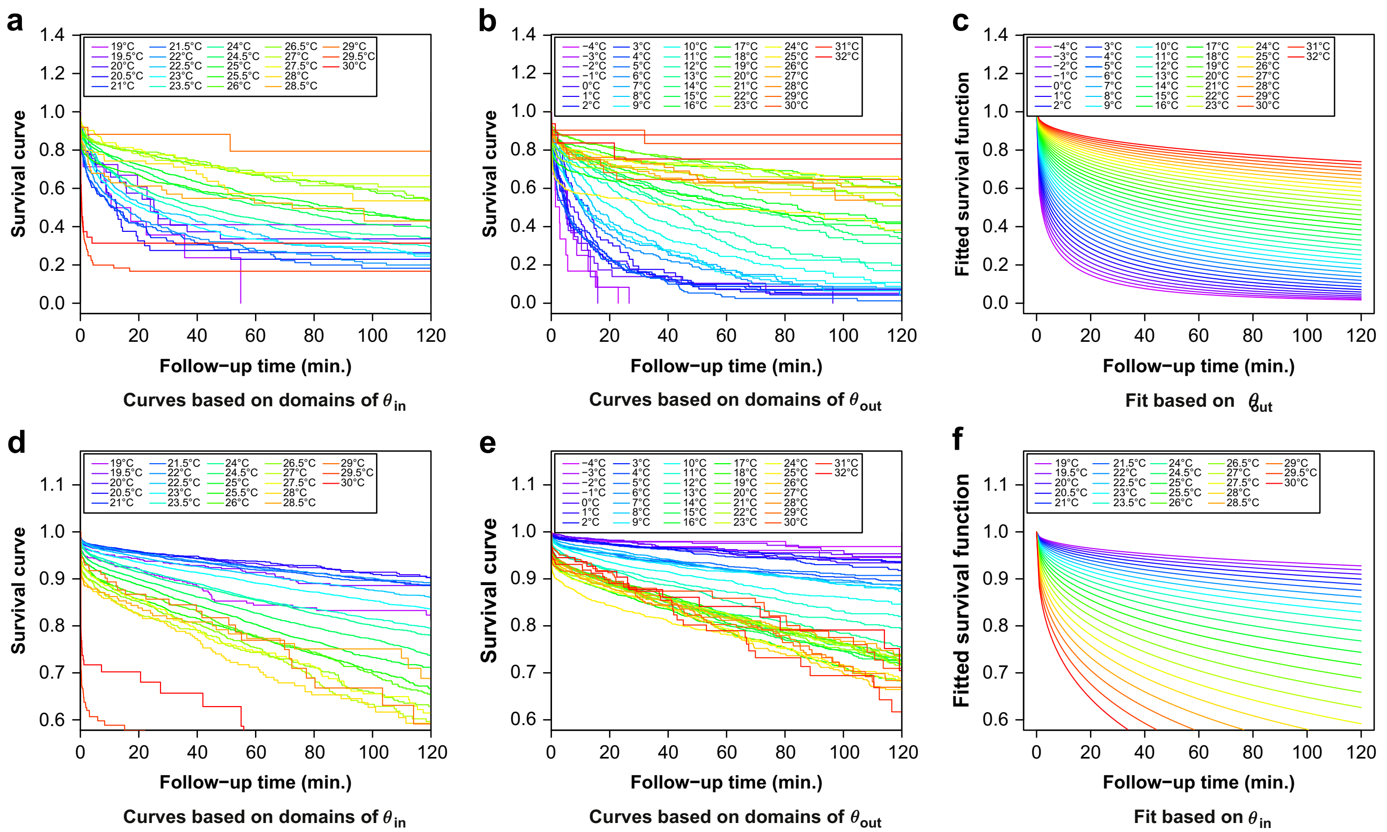


Fig. 6. Kaplan–Meier estimators of survival functions of opening (top) and closing (bottom) duration, where points show censored data (left and middle). Fitted parametric survival functions are shown in right.

(Fig. 6(e)). Conversely, the survival curves vary clearly for different values of θ_{in}. Furthermore, we can straightforwardly interpret the immediate decays along the ordinates in closing durations as *opening probabilities on arrival*, that increase strongly with θ_{in} as expected. Intermediate openings are then described by the rest of the curve, with higher proportional decays being observed for higher temperature.

As for openings, we once again use the Weibull distribution to describe closing durations. We include first θ_{in}, and notice that the addition of θ_{out} is significant. Based on 183,949 time intervals with closed windows, we obtain for the shape and the scale $\log(1/\alpha) = 0.866 \pm 0.00635$, while the scale is $\lambda = 1/\exp((15.789 \pm 0.28596) + (-0.249 \pm 0.01279) \cdot \theta_{in} + (-0.105 \pm 0.00312) \cdot \theta_{out})$, with $R^2 = 0.033$. For illustration purposes, we show the fitted survival function with θ_{in} as the driving variable in Fig. 6(f).

4.4.3. Remarks

The main limitation in the modelling of actions with survival curves is the obvious risk in predicting in advance potentially long opening times, independently of subsequent variations of environmental stimuli. For instance, as a transition from closed to open is performed, the indoor conditions evolve in response to heat transfers with the outdoors.

We observe that opening and closing durations are poorly fitted by an exponential distribution. The memoryless property referenced above that defines discrete-time Markov processes is thus not fulfilled, which suggests that in this context window opening and closing processes are not memoryless, and so are not fully appropriately modelled by a Markov process.

The obtained Weibull distributions confirm once again that delayed opening of windows is mainly caused by indoor stimuli, while the main driving stimulus for window closings is external (the feedback of the opening).

4.5. Integration of individual behaviours

The models developed above were derived from data relating to the whole set of occupants and for the entire surveyed period. We examine here variations in behaviour among the surveyed occupants and provide a method to account for the observed behavioural diversity.

4.5.1. Variability between occupants

We provide in Table 6 the principal indicators concerning general conditions and behavioural differences for all the surveyed occupants[10] (or their combinations). In order to distinguish between "active" and "passive" occupants we use as an indicator the percentage of time occupants leave their windows open during occupancy. We see that this ratio ranges from 13.3% to 56.4%, with half of the occupants grouped between 20% and 40%. Another possible and related indicator is the mean amount of actions per unit occupancy time and per occupant, which offers a similar activity hierarchy between occupants, and directly considers the frequency of actions taken. In the last column of Table 6 we show

[10] Occupants with total presence duration inferior to 250 days are not included here, as their behaviour could not be observed for a fully representative period, which may induce biased indicators. The data used in the previous sections thus do not exactly correspond to the aggregated data of individual occupants presented in Table 6.

Table 6
Variability between occupants: reference, number of occupants, survey duration, distance to closest window, overall proportion of time open, actions per person and per week, indoor and outdoor characteristic temperatures, presence of predictive behaviour and overall activity index.

Ref.	Nb. Pers.	Survey [days]	Dist. [m]	Ratio open	Actions	$\theta_{in,50}$ [°C]	$\theta_{out,50}$ [°C]	Predictive	Overall activity
001-D1	2	730	1.5	21.0%	1.33	26.8	22.7	No	Average
001-D2	2	451	1.5	56.4%	4.62	20.4	6.9	No	High
002-D1	2	275	1.5	40.4%	5.85	23.7	17.2	No	High
002-D2	2	334	1.5	45.2%	2.40	27.8	19.0	No	Average
003-S1	1	2222	1.5	14.1%	0.81	27.0	28.0	No	Low
004-D1	2	321	1.5	18.1%	0.87	32.4	18.5	Yes	Low
004-S1	1	1706	1.5	29.8%	2.40	24.0	18.4	No	Average
101-D1	2	606	NA	18.8%	1.59	27.6	22.8	No	Average
101-S1	1	1310	3	16.5%	1.13	26.7	21.0	No	Low
103-S1	1	272	2.2	13.3%	0.41	27.5	24.9	No	Low
103-S2	1	1949	2.2	33.7%	1.52	18.3	28.1	Yes	Average
104-S1	1	2222	5	21.5%	0.68	NA	NA	Yes	Low
105-S1	1	2222	3.2	38.3%	4.91	25.2	16.0	No	High
106-D1	2	2222	2.8	28.5%	1.55	26.6	18.6	No	Average
201-S1	1	2222	3	43.3%	3.24	23.0	14.9	Yes	Average
203-D1	1	1550	1.5	24.6%	1.22	28.1	24.7	NA	Average
204-S1	1	1048	1.5	36.0%	3.91	24.9	17.2	Yes	High
204-S2	1	1173	1.5	46.7%	4.02	24.3	12.3	No	High
Single (10)	1			29.3%	2.30	24.5	20.1		
Double (8)	2			31.6%	2.43	26.7	18.8		

the results of a possible classification: occupants whose average amount of actions per unit time lies in the interquartile range are classified as "average", while others are placed in the categories "low" and "high" overall activity.

Furthermore, different occupants may vary not only in the intensity of their behaviour but also in its nature. For example in Fig. 7 we present logit probability distributions observed for different occupants. It can be seen that a minority of occupants is weakly influenced by thermal stimuli. Nevertheless, for the majority of them, the slopes of the univariate logit distributions for θ_{in} and θ_{out} are not significantly different; the differences arising mainly in the intercepts. Variability may thus be meaningfully summarised by the characteristic temperature θ_{50}. A weak relationship may be noticed between $\theta_{in,50}$ and the number of actions per week.

It is worth noticing that some occupants have open windows at high θ_{out}, while others follow the decreasing trend shown in Fig. 3(b) to prevent the incoming of hot outside air. We qualify this behaviour as "predictive" if the polynomial terms in the logit are significant (see Table 6). We see that five observed occupants adopt this preventive strategy.

Similarly, differences in behavioural patterns between occupants may be noticed if we derive individual transition probabilities for the discrete-time Markov process. These results suggest that occupants generally display the same type of behaviour, but at higher or lower temperatures. Furthermore, the predictive scheme described above, i.e. refraining from opening windows for hot outside conditions, is reproduced by the same occupants.

4.5.2. Group actions

No obvious difference in behaviour related to total opening duration is distinguishable in Table 6 between offices with one or two occupants. In this latter case, one possibility is to assume that occupants act independently and the "activity" in an office is aligned to that of the most active (or assertive) of the occupants. Similarly, none of the distributions displayed in Fig. 7 show differentiated behaviour between single and double-occupied offices.

When using the discrete- or continuous-time random processes to predict group actions, two variants may be suggested. A first possibility is to explicitly model the occupancy of each potential occupant, and launch the window opening model for each of them. An alternative

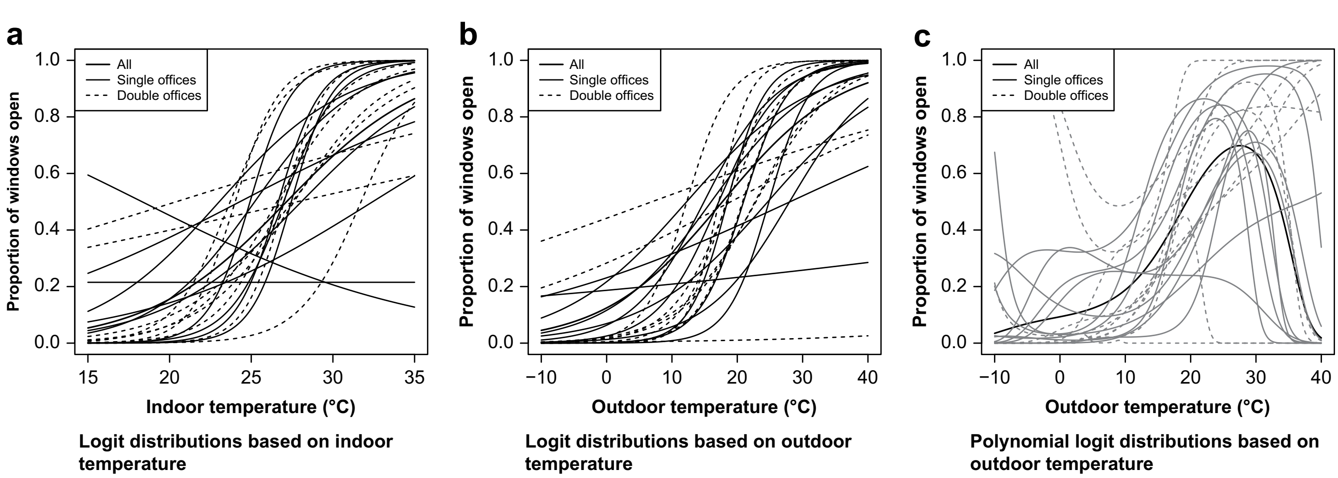


Fig. 7. Occupant specific probability distributions.

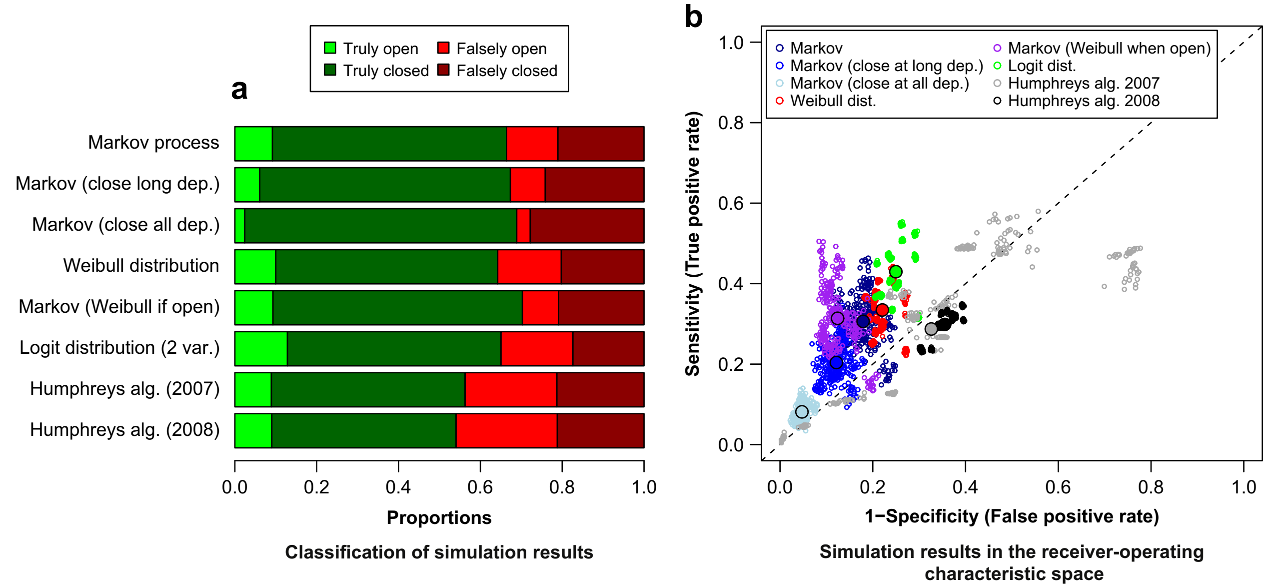


Fig. 8. Cross-validation results.

would be to adapt the sub-model for actions during presence to the special case where a second arrival may occur in an already occupied office. This sub-model will then predict higher transition probabilities.

5. Discussion

5.1. Summary

We observed that indoor conditions describe opening actions better than do outdoor conditions – this being our interaction stimulus. But closing actions tend to be better described by outdoor conditions, based on perceived draughts or a risk of overheating when $\theta_{out} > \theta_{in}$; likewise whether windows will be left open overnight for cooling purposes. Therefore if we consider the aggregate dataset it is understandable that for a univariate probability distribution θ_{out} is statistically stronger than θ_{in}, but this does not make it a better model. This is partly because the previously mentioned subtleties are ignored and partly because, as noted earlier, when using θ_{out} alone the predicted window states are independent of the design of the building; so that occupants of very different adjacent buildings (e.g. with minimal and high façade glazing ratio) would be predicted to interact with their windows with similar probability. For such a hypothetical building θ_{out} may again be the best predictor for the (aggregate) logit distribution, but with drastically different parameters; in particular $\theta_{out,50}$ will be lower in the highly glazed case.

Any distribution based on θ_{out} is thus strongly building-dependent and without generality, requiring separate calibration for each building to which it is applied – an impossible task. Furthermore, it cannot be excluded that the distribution with both θ_{in} and θ_{out} could be dampened by this lack of generality. On the other hand, although the logit distribution with just θ_{in} is expected to describe occupants' actions with more generality, this also misses important subtleties, particularly in relation to the closing of windows, which undermine its predictive accuracy.

The obtained transition probabilities for the discrete-time Markov process solve this problem, as they directly link the probability for an occupant to take action with the direct environmental stimulus (θ_{in}), whilst also accounting for the fact that θ_{out} has a determinant influence on intermediate closing probability (the sole situation where θ_{out} has a direct effect on the occupant). Transition probabilities have also been derived which account for variations due to individual behaviours (Section 4.5). A possible lack of generality is that the closing probabilities which depend on θ_{out} are likely to depend on window size and opening angle, possibly needing further calibration according to these parameters. The same remarks apply to our continuous-time random process.

In summary then not only do the presented models improve the quality of predictions; they also account for the real stimuli motivating adaptive actions, so improving upon their generality.

5.2. Cross-validation

In this paper we have presented models of occupants' interactions with windows based on three different methods. We describe here our validation procedure to perform a consistent evaluation of their predictive powers. In addition to this, we will also compare the results from these models with a Bernoulli random variable with constant probability (a random guess based on observed overall opening proportion), with previous published work (the two versions of the Humphreys algorithm) and with variants based on our models (eg. a discrete-time Markov process with partial sub-models for departure[11] and a hybrid model).

We assess the predictive power of the models by checking four aspects:

- Discrimination: Does the model reproduce well the list of observed window states?
- Overall prediction: Does the model predict a consistent overall fraction of openings throughout the simulation period?
- Dynamics: Does the model predict consistent delays between actions?
- Aggregated results: Is the predicted total number of open windows consistent with observations?

Based on these criteria we will retain the best performing model.

For this validation exercise we have performed 20 repeated simulations using 5-min time steps for the whole period with available measurements for the 14 measured offices, producing

[11] This allows a comparison with the approach taken by Yun and Steemers [21,22].

Table 7
Validation parameters: true positive rate, false positive rate, accuracy, total proportion of simulated time steps with window open, median duration (min.) of openings and closings, median error and interquartile range of error on total number of windows open.

Model	TPR	FPR	ACC	Prop. open	Dur. open	Dur. closed	Error (median)	Error (IQR)
Exact	100.0%	0.0%	100.0%	30.2%	110	1290	0.000	0.000
Logit dist. (with θ_{in})	33.4%	29.5%	59.1%	30.6%	5	10	+2.714	3.286
Logit dist. (with θ_{out})	42.9%	25.3%	64.8%	30.5%	5	10	+1.571	2.000
Logit dist. (with θ_{in} and θ_{out})	43.0%	25.0%	65.1%	30.4%	5	10	+1.381	2.048
Logit dist. (with polyn. of θ_{out})	43.7%	24.4%	65.5%	30.2%	5	10	+1.571	1.619
Markov	30.6%	17.9%	66.4%	21.8%	710	2895	+0.714	1.667
Markov (close at long dep.)	20.4%	12.2%	67.3%	14.6%	205	4090	−0.143	2.905
Markov (close at all dep.)	8.1%	4.7%	68.9%	5.7%	40	3150	−0.952	4.095
Weibull distribution	33.4%	22.1%	64.2%	25.5%	5	5	+2.286	2.714
Markov (Weibull for openings)	31.4%	12.4%	70.3%	18.1%	250	2975	+0.286	1.190
Humphreys algorithm (2007)	28.7%	32.6%	56.3%	31.4%	12820	20480	+1.667	6.286
Humphreys algorithm (2008)	60.3%	76.5%	34.7%	71.5%	5	10	+3.476	5.714
Random guess	30.2%	30.2%	58.6%	30.2%	5	10	+2.667	4.381

$20 \times 14 = 280$ sets of simulated window states $W_{sim}(t)$, to be compared with 14 sets of observed states $W_{obs}(t)$. This procedure was repeated for each of our models as well as for the Humphreys algorithm.[12]

5.2.1. Discrimination

A general validation procedure should involve comparing each model's ability to directly reproduce observed window states. We would thus obtain results that may be classified in four groups: a predicted positive outcome is (i) truly positive (*TP*), (ii) falsely positive (*FP*, Type I error); a predicted negative outcome is (iii) truly negative (*TN*), (iv) falsely negative (*FN*, Type II error). We present these proportions for each simulated model in Fig. 8(a).

We may then accumulate these results to define the overall *true positive rate* (or *sensitivity*, or *hit rate*) $TPR = TP/(TP + FN)$, the *false positive rate* (or *fall-out*) $FPR = FP/(FP + TN)$ and the *specificity* $SPC = 1 - FPR$. We may then also compute the *accuracy* defined as $ACC = (TP + TN)/(P + N)$, which gives the proportion of correct predictions.

Based on the twenty repeated simulations, these indicators are computed and displayed in Table 7. Applying the concepts introduced in Appendix A, we may also draw the corresponding points in the receiver-operating space (Fig. 8(b)), in which our ideal model would be located at minimum x and maximum y. Each small point corresponds to a single simulation of an office, while parameters referring to the aggregated results of a model are plotted as bigger solid points.

As expected from our previous statistical tests, a Bernoulli random variable based on the univariate logit distribution with θ_{in} performs much worse than with θ_{out}. The distribution with two variables discriminates slightly better compared with θ_{out} alone, while the polynomial logit with θ_{out} offers best discrimination among the logit distributions.

The discrete-time Markov process gives lower values of *TPR* and *FPR*. This suggests that this model is more "conservative", that is it predicts less openings than the logit distributions (and misses slightly more of them), but on the other hand much less false openings are predicted. Furthermore, it has higher overall accuracy. We can also observe that the quality of predictions decreases drastically if we attempt to treat actions on departure simplistically.

The Weibull distribution for the continuous-time random process is generally slightly less "conservative" than the Markov process, albeit with a slightly lower accuracy. However, this model is computationally much faster. In an attempt to find a good compromise between accuracy and speed we have therefore developed a hybrid model based on the discrete-time Markov process, using a Weibull distribution for opening durations only. Our simulations show that this provides the highest accuracy, while increasing *TPR* and reducing *FPR* compared with the plain Markov model. According to the discrimination criteria, this model offers the best performance.

5.2.2. Overall window opening ratio

Based on the total presence duration $T_{pres,tot}$ and the total window opening time $T_{open,tot}$, we define for each office the overall window opening ratio as $r_{open} = T_{open,tot}/T_{pres,tot}$. We show observed and predicted values in Table 7. Overall opening ratios predicted by the Markov processes are rather low, particularly when night ventilation behaviour is neglected. The logit distributions generally reproduce well this parameter, likewise the Humphreys algorithm. However the Markov process predicts the greatest variability between results.

It is worth noting that all the models predict similar total opening ratios between simulated offices, which show that the added refinements do not reproduce this variability between occupants.

5.2.3. Opening and closing median durations

As expected (Table 7), the models based on Bernoulli random variables do not predict coherent delays between actions, as they are not explicitly based on any dynamics in their formulation. The Markov model overestimates these durations; that is it does not predict enough actions from occupants. Not surprisingly, the Weibull distribution best reproduces delays between actions, as they

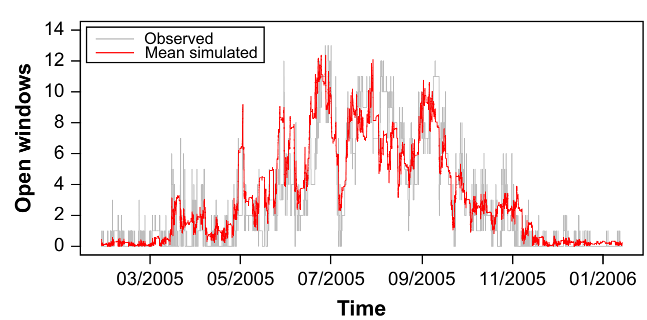


Fig. 9. Observed and mean simulated number of windows open on a period of a year.

[12] We used a probability distribution adjusted to our data to simulate the first version of the Humphreys algorithm (2007) and the original published parameters for its second version (2008).

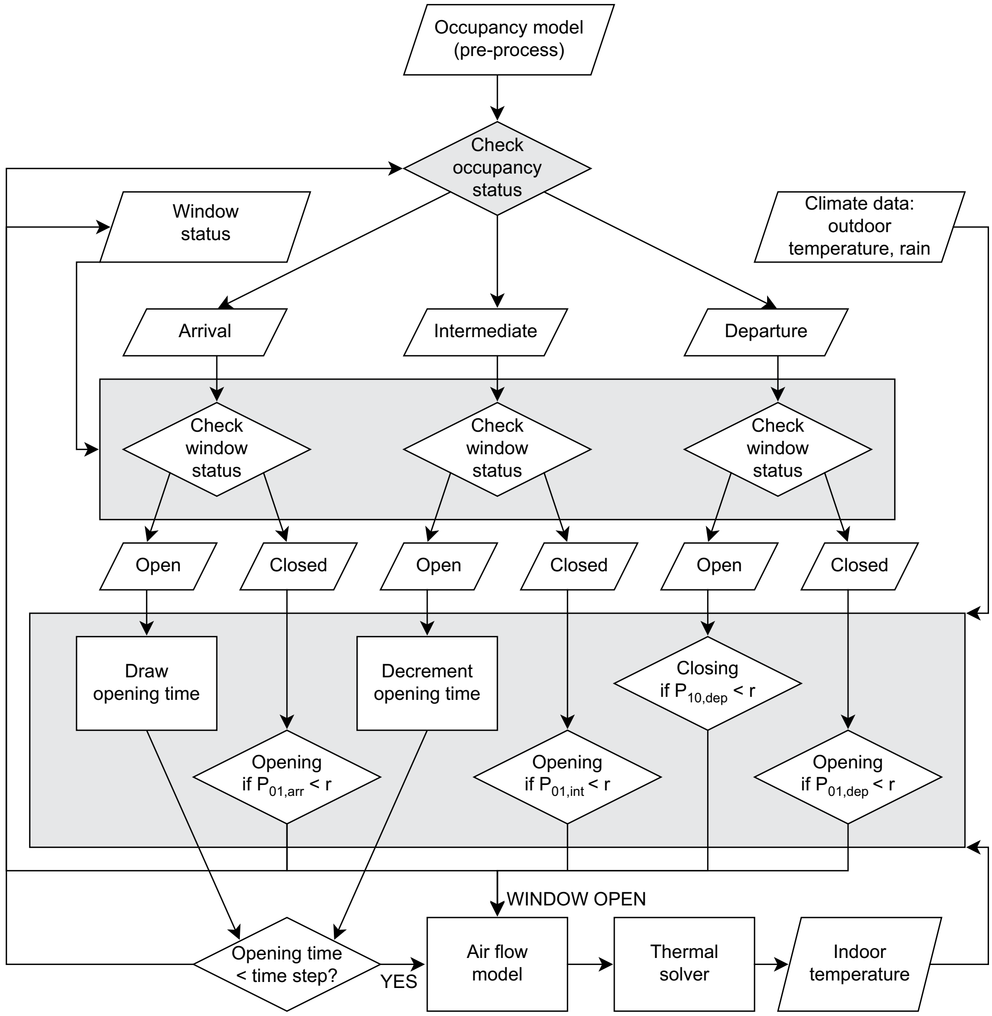


Fig. 10. Implementation scheme of the window opening algorithm.

are directly based on observed opening and closing durations, while the hybridised model combines results from these latter models. As with the Markov model, the Humphreys algorithm (2007) predicts durations that are too long, which may be caused by too large a deadband or by ignoring occupancy transitions.

5.2.4. Aggregated results

We extend here our validation approach to the global simulation results and check whether the models are able to reproduce observed window openings on a large scale. Based on individual simulations of window states in the 14 offices, we compute the total number of windows open $N_{\text{sim}}(t) = \sum_{k=1}^{14} W_k(t)$ for each simulation. Averaging over the 20 sets of $N_{\text{sim}}(t)$ we deduce the mean predicted number of windows open at each time step $\overline{N}_{\text{sim}}(t)$. Comparing with the observed number of windows open $N_{\text{obs}}(t)$, we may compute the error $\varepsilon(t) = N_{\text{sim}}(t) - N_{\text{obs}}(t)$ at each time step.

We show in Table 7 the median and interquartile range of this error for each model. From this we observe that the hybrid model offers the best performance on aggregated results, with smallest error magnitude. Furthermore, comparisons of measured and predicted number of windows open (Fig. 9) indicate that this model reproduces well the temporal variation of window openings.[13]

[13] This refers to the period from 27 January 2005 to 14 January 2006, which offers representative climatic conditions and uninterrupted measurements.

5.2.5. Overall recommendation

From these validation results, we recommend the use of a hybridised model including a discrete-time Markov model for the prediction of openings and a Weibull distribution for their duration. Although, this model somewhat underestimates the overall opening ratio, we observed that it offers the highest accuracy, produces the best discrimination between window states, reproduces acceptably the delays between actions and offers the most reliable aggregated predictions at the scale of a whole building. A detailed description of this algorithm for implementation in building simulation tools is provided in Appendix B.

6. Conclusion

Based on almost seven years of observations we have developed three different modelling methods for the prediction of actions on windows: a Bernoulli process based on a logit probability distribution, a discrete-time Markov process with sub-models for different occupancy statuses and extended this latter to a continuous-time random process. Supported by rigorous cross-validation, we have demonstrated the superiority of a discrete-time Markov process approach and its strong added value compared with existing models. We have furthermore inferred a continuous-time model that could be efficiently used for a fast calculation of opening and closing durations.

We have finally tested possible combinations in these approaches and selected a hybrid model. This hybrid combines the accuracy of the discrete-time Markov process with the efficiency of the continuous-time model for opening durations. For this we also describe a step-by-step process by which the algorithm may be implemented.

We have also studied the diversity in individual behaviours and described a possible method to integrate them if necessary; likewise a method to integrate in the Markov model the possibility of acting on multiple windows.[14]

However, there remain some outstanding issues to be addressed in the modelling of window opening and closing behaviour; particularly in respect to the angle of window opening as opposed to simply the state open and closed. Finally we are mindful that, although we have demonstrated the validity of the formulation of our proposed new algorithm, the parameters for its calibration are strictly speaking currently limited to just one building type.

It would thus be desirable to make use of measurements from other buildings (residential in particular), in which opening angles are also recorded, to have a stronger basis for calibration and thus application to other simulated buildings. Such surveys might also usefully include other variables which may influence actions on windows, such as radiant temperature or indoor relative humidity (particularly for tropical climates). Factors related to indoor air quality (eg. CO_2 or pollutant concentration) should also be treated; however it is plausible that the inclusion of T_{pres} in our intermediate openings model ($P_{01,int}$) could implicitly account for this (at least in part).

Acknowledgments

Financial support received from the European Commission as part of the *CONCERTO II Project HOLISTIC* is gratefully acknowledged. We warmly thank the present and former staff of our laboratory who contributed to the installation and maintenance of the data acquisition sensors, particularly Antoine Guillemin, David Lindelöf and Laurent Deschamps.

A. Statistical tests for logistic models

A.1. Assessment of statistical significance

Likelihood ratio test statistic: We define the *deviance* of a generalised linear model with associated likelihood $l_{\text{fitted}}(\beta)$ as

$$D = -2\log\left(\frac{l_{\text{fitted}}(\beta)}{l_{\text{sat}}(\beta)}\right) = -2\log\left(l_{\text{fitted}}(\beta)\right), \tag{13}$$

with $l_{\text{sat}}(\beta)$ being the likelihood of a saturated model, which may be shown to be equal to one. The deviance may be interpreted similarly as the residual sum-of-squares in linear regression. We may now define a test for the significance of k independent variables added to a given model. Let us define a model without the variable with likelihood l_0 and deviance D_0, and another model with l_1 and D_1. The *likelihood ratio test statistic* or *G-statistic* is defined as the deviance difference caused by the inclusion of k predictors,

$$G = D_0 - D_1 = -2\log\left(\frac{l_0(\beta)}{l_1(\beta)}\right), \tag{14}$$

and has a χ^2 distribution with k degrees of freedom. The associated p-value allows us to check the significance of adding new variables to the model.

Wald test: The Wald test for the significance of an independent variable is obtained by dividing the estimate of the obtained slope parameter $\hat{\beta}$ by its standard error $\text{SE}(\hat{\beta})$, giving $W = \hat{\beta}/\text{SE}(\hat{\beta})$. Under the null hypothesis, this ratio follows a standard normal distribution, and the two tailed p-value $P(|z| > W)$ where $z \sim N(0,1)$ is easily obtained. This test should not be used indiscriminately, as previous research [33] has shown that it often fails to reject the null hypothesis for significant predictors.

A.2. Measures of goodness-of-fit

Statistical significance itself does not directly provide for an assessment of the quality of predictions given by a logit distribution. Furthermore, when a large database is used, a factor with tiny predicted effect may be statistically significant, but will not bring very useful improvements in response prediction. Several criteria have therefore been developed to give an idea of how well a given distribution predicts the original outcome variable, i.e. the goodness-of-fit of the model. In this paper we present the area under the ROC curve (*AUC*), the Nagelkerke R^2, the Brier score B and Somers' D_{xy}.

Area under ROC curve: Having defined the probability of finding the outcome variable in the positive state $P(Y = 1|x)$, we use a cutpoint c and set $Y=1$ if $P(Y = 1|x) > c$ and $Y=0$ for $P(Y = 1|x) \leq c$. Comparing values of Y predicted through this cutpoint with observed values of Y, we may encounter the four cases described in Section 5.2 (truly or falsely positive or negative) and compute the associated sensitivity and specificity.

We may plot the sensitivity against the complementary of the specificity for different values of the cutpoint between 0 and 1, giving the *receiver operating characteristic (ROC) curve* [34]. The area under the ROC curve will be called here the *AUC* index, and is a direct measure of the discriminating power of a given model. It may take values between 0.5 (we may as well toss a coin) and 1.0 (exact predictions), but values above 0.7 are generally considered as acceptable discrimination [29].

Nagelkerke's generalised R^2: Nagelkerke's generalised R^2 measures the proportion of explained deviance in a model and is defined as

$$R^2 = \frac{1 - \exp((D - D_{\text{null}})/n)}{1 - \exp((-D_{\text{null}})/n)}, \tag{15}$$

where D and D_{null} are the deviances of the considered model and the null model, based on n observations. It extends the well-known definition of the proportion of explained variance R^2 used in linear models to GLMs and particularly logistic models. Values of R^2 obtained with this convention are generally much lower than their linear model counterparts.

Brier score: The Brier score measures the accuracy of a set of probability assessments. It is defined as the mean value of the

[14] For n windows, transition probabilities $P_{ij}(\mathbf{x})$ from i to j windows open ($0 \leq i, j \leq n$) are arranged in an $(n+1)\times(n+1)$ matrix. For instance, with three simulated windows and two currently open, the closing of one or both of open windows is modelled by P_{21} or P_{20} in the case of both and a further opening by P_{23}, with $P_{22} = 1 - P_{20} - P_{21} - P_{23}$. These transition's probabilities may be included in a 4×4 matrix containing 12 independent elements:

$$P_{ij} = \begin{pmatrix} P_{00} = 1 - \sum_{k\neq0} P_{0k} & P_{01} & P_{02} & P_{03} \\ P_{10} & P_{11} = 1 - \sum_{k\neq1} P_{1k} & P_{12} & P_{13} \\ P_{20} & P_{21} & P_{22} = \dots & P_{23} \\ P_{30} & P_{31} & P_{32} & P_{33} = \dots \end{pmatrix}.$$

This requires the inference of $n(n+1)$ probability distributions for each occupancy transition. We may similarly derive survival curves for delays until actions on additional windows.

squared difference between observed outcomes and their predicted probabilities:

$$B = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2, \tag{16}$$

where P_i are the predicted probabilities and Y_i the corresponding observed values for the outcome variable, for each of n observations. It follows directly from the definition that a lower Brier score indicates a higher accuracy.

Somers' D_{xy} rank correlation: Somers' D_{xy} parameter is defined as the difference between concordance and discordance probabilities. Given two individual values of the predictors x_0 and x_1, randomly sampled from the populations with outcome variables $Y_0 = 0$ and $Y_1 = 1$, respectively. We define then $D_{xy} = P(Y_1 > Y_0) - P(Y_0 > Y_1)$. Somers' D_{xy} may take values between 0 (random predictions) and 1 (perfect discrimination).

B. Integration into building simulation tools

All the models for the prediction of window openings presented here may be integrated in any dynamic simulation environment. We present here the detailed steps for the implementation of our discrete-time Markov process hybridised with a Weibull distribution to predict opening durations. We assume that climate data are available and that occupancy is first predicted through a pre-processor for the whole simulation period. A general scheme describing the implementation procedure for a dynamic simulation of window states $W(t)$ for time steps of length $\delta t = (t_{i+1} - t_i)$ is provided in Fig. 10, which consists of the following steps:

1. The occupancy status (absence, arrival, ongoing presence or departure) is retrieved, together with the concomitant presence or absence durations. The variables $\theta_{\mathrm{out}}(t_i)$, $\theta_{\mathrm{out,dm}}(t_i)$ and $f_{\mathrm{R}}(t_i)$ are obtained from the climate data and $\theta_{\mathrm{in}}(t_i)$ from a coupled thermal solver.
2. **Case 1.** If the occupant is absent, the window state is set as identical to its previous state: $W(t_i) = W(t_{i-1})$.
2. **Case 2.** If the occupant arrives and the window is closed ($W(t_{i-1}) = 0$):
   a) We compute $p(t_i) = P_{01,\mathrm{arr}}(t_i)$ and draw a random number $0 \le r_i \le 1$ from a uniform distribution.
   b1) If $p(t_i) > r_i$ we set $W(t_i) = 0$.
   b2) If $p(t_i) \le r_i$, we draw a random number d_i from the Weibull distribution $f_{\mathrm{o}}(\theta_{\mathrm{out}})$. If $d_i < \delta t$, we set $W(t_i) = 0$. If $d_i \ge \delta t$, we set $W(t_i) = 1$ and $d_{i+1} = d_i - \delta t$.
2. **Case 3.** If the occupant arrives and the window is open ($W(t_{i-1}) = 1$), we draw a random number d_i from the Weibull distribution $f_{\mathrm{o}}(\theta_{\mathrm{out}})$, and perform the same procedure as in Case 2 (b2).
2. **Case 4.** If the occupant is in intermediate presence and the window is closed ($W(t_{i-1}) = 0$):
   1. We compute $P_{01,\mathrm{int}}(t_i)$ and draw a random number $0 \le r_i \le 1$ from a uniform distribution.
   2. If $p(t_i) > r_i$ we set $W(t_i) = 0$.
   3. If $p(t_i) \le r_i$, we draw a random number d_i from the Weibull distribution $f_{\mathrm{o}}(\theta_{\mathrm{out}})$, and perform the same procedure as in Case 2 (b2).
2. **Case 5.** If the occupant is in intermediate presence and the window is open ($W(t_{i-1}) = 1$), we retrieve the decremented opening duration d_i calculated at the previous opening. If $d_i < \delta t$, we set $W(t_i) = 0$. If $d_i \ge \delta t$, we set $W(t_i) = 1$ and $d_{i+1} = d_i - \delta t$.
2. **Case 6.** If the occupant leaves:
   a) We compute $p(t_i) = P_{01,\mathrm{dep}}(t_i)$ if $W(t_{i-1}) = 0$ (or $p(t_i) = P_{10,\mathrm{dep}}(t_i)$ if $W(t_{i-1}) = 1$) and draw a random number $0 \le r_i \le 1$ from a uniform distribution.
   b1) If $p(t_i) > r_i$ we set $W(t_i) = 0$ (or $W(t_i) = 1$).
   b2) If $p(t_i) \le r_i$, we set $W(t_i) = 1$ (or $W(t_i) = 0$).
3. The volume of air exchanged is computed, the thermal solver predicts $\theta_{\mathrm{in}}(t_{i+1})$ and we start the next time step.

References

[1] Dick JB, Thomas DA. Ventilation research in occupied houses. Journal of the Institution of Heating and Ventilating Engineers 1951;19:306–26.
[2] Brundrett GW. Ventilation: a behavioural approach. Energy Research 1977;1:289–98.
[3] Lyberg MD. Energy losses due to airing by occupants. In: CIB commission W67 symposium – energy conservation in the built environment, Dublin, Ireland; 1982.
[4] Warren PR, Perkins LM. Window-opening behavior in office buildings. ASHRAE Transactions 1984;90(1B):1056–76.
[5] Fritsch R, Kohler A, Nygård-Ferguson M, Scartezzini JL. A stochastic model of user behaviour regarding ventilation. Building and Environment 1990;25(2):173–81.
[6] Roulet CA, Cretton P, Fritsch R, Scartezzini JL. Stochastic model of inhabitant behavior with regard to ventilation. Technical report; 1991.
[8] Nicol JF, Roaf SC. Pioneering new indoor temperature standards: the Pakistan project. Energy and Buildings 1996;23:169–74.
[9] Nicol JF, Raja IA, Allaudin A, Jamy GN. Climatic variations in comfortable temperatures: the Pakistan projects. Energy and Buildings 1999;30(3):261–79.
[10] Raja IA, Nicol JF, McCartney KJ. Natural ventilated buildings: use of controls for changing indoor climate. Renewable Energy 1998;15:391–4.
[11] Raja A, Nicol JF, McCartney KJ, Humphreys MA. Thermal comfort: use of controls in naturally ventilated buildings. Energy and Buildings 2001;33(3):235–44.
[12] McCartney KJ, Nicol JF. Developing an adaptive control algorithm for Europe. Energy and Buildings 2002;34(6):623–35.
[13] Nicol JF. Characterising occupant behaviour in buildings: towards a stochastic model of occupant use of windows, lights, blinds, heaters and fans. In: Seventh international IBPSA conference proceedings, Rio de Janeiro; 2001.
[14] Nicol JF, Humphreys MA. A stochastic approach to thermal comfort – occupant behaviour and energy use in buildings. ASHRAE Transactions 2004;110(2):554–68.
[15] Robinson D. Some trends and research needs in energy and comfort prediction. In: Windsor conference 2006; 2006.
[16] Rijal HB, Tuohy P, Humphreys MA, Nicol JF, Samuel A, Clarke J. Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings. Energy and Buildings 2007;39(7):823–36.
[17] Rijal HB, Tuohy P, Nicol JF, Humphreys MA, Samuel A, Clarke J. Development of an adaptive window-opening algorithm to predict the thermal comfort, energy use and overheating in buildings. Journal of Building Performance Simulation 2008;1(1):17–30.
[18] Robinson D, Campbell N, Gaiser W, Kabel K, Le-Mouel A, Morel N, et al. Suntool – a new modelling paradigm for simulating and optimising urban sustainability. Solar Energy 2007;81(9):1196–211.
[19] Fanger PO. Introduction of the olf and the decipol units to quantify air pollution perceived by humans indoors and outdoors. Energy and Buildings 1988;12(1):1–6.
[20] Haldi F, Robinson D. On the behaviour and adaptation of office occupants. Building and Environment 2008;43(12):2163–77.
[21] Yun GY, Steemers K. User behaviour of window control in offices during summer and winter. In: CISBAT, Lausanne, Switzerland; 2007.
[22] Yun GY, Steemers K. Time-dependent occupant behaviour models of window control in summer. Building and Environment 2008;43(9):1471–82.
[23] Rijal HB, Tuohy P, Humphreys MA, Nicol JF, Samuel A, Raja IA, et al. Development of adaptive algorithms for the operation of windows, fans and doors to predict thermal comfort and energy use in Pakistani buildings. ASHRAE Transactions 2008;114(2):555–73.
[24] Herkel S, Knapp U, Pfafferott J. A preliminary model of user behaviour regarding the manual control of windows in office buildings. In: Ninth international IBPSA conference on building simulation 2005, Montréal, Canada; 2005.
[25] Herkel S, Knapp U, Pfafferott J. Towards a model of user behaviour regarding the manual control of windows in office buildings. Building and Environment 2008;43(4):588–600.
[26] Reinhart C. Lighswitch-2002: a model for manual and automated control of electric lighting and blinds. Solar Energy 2004;77:15–28.
[27] Andersen RV, Toftum J, Andersen KK, Olesen BW. Survey of occupant behaviour and control of indoor environment in Danish dwellings. Energy and Buildings 2009;41(1):11–6.
[28] Dobson AJ. An introduction to generalized linear models. 2nd ed. Chapman & Hall; 2002.
[29] Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York, USA: John Wiley & Sons; 2000.
[30] D.G. Kleinbaum and M. Klein. Survival analysis – a self-learning text. Springer; 2005.

[31] R Development Core Team. R: a language and environment for statistical computing; 2007. ISBN 3-900051-07-0.
[32] de Dear RJ, Brager GS. Thermal comfort in naturally ventilated buildings: revisions to ASHRAE standard 55. Energy and Buildings 2002;34(6):549–61.
[33] Hauck WW, Donner A. Wald's test as applied to hypothese in logit analyses. Journal of the American Statistical Association 1977;72:851–3.
[34] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006;27(8):882–91.
[35] Yun GY, Tuohy P, Steemers K. Thermal performance of a naturally ventilated building using a combined algorithm of probabilistic occupant behaviour and deterministic heat and mass balance models. Energy and Buildings 2009;41(5):489–99.