University of Liège
School of Engineering

Consolidation Milestone

PROJ0016: Big Data Project

Julien Gustin, Romain Charles, Joachim Houyon

December 7, 2021

1 Nomenclature

T_z :	Air temperature of each zone	[°C]	(1)
T_a :	Ambient temperature (outdoor)	[°C]	(2)
T_w :	Water temperature	[°C]	(3)
$T_{s,z}$:	Temperature set	[°C]	(4)
Δt :	Time step length (5min)	$[s]$	(5)
t :	Time step	$[s]$	(6)
R_z :	thermal resistance (for thermal loss)	[°C/W]	(7)
$R_{z,n}$:	Thermal resistors coupling the zones z and n	[°C/W]	(8)
C_z :	thermal capacities (Heat capacity of the building)	[J/°C]	(9)
m_z :	Mass flow rate	[kg/s]	(10)
k_z :	Thermal loss coefficient	[%]	(11)
Q_{heat} :	Heat transfer from the radiator to the building	$[W]$	(12)
cp :	Heat capacity of water (4810)	[kJ/kg/°C])	(13)
a_z :	Offset parameter - can be seen as internal heat gains		(14)

$\forall z \in$ zones and $\forall n \in$ neighbour(z)

2 Description of the data-set

The data we have are room temperatures, set temperatures for each room, outdoor temperature, heat system temperature and PV output. All these data were measured at a time t with Δt= 5 minutes. They are therefore provided with the corresponding time series.

Most of the data provided for this work was used. The room temperatures and the outdoor temperature were used to evaluate the resistance and heat capacity components of the rooms. But also the set temperatures and the water temperatures to establish the influence that the radiator will have.

We could have used the PV output, which was provided, to include solar input or room measurements, for example, to make our calculations even more accurate.

By looking at the data set we have seen that in winter, living room has some strange behavior such as having high indoor temperature while it is cold outside and temperature set is low, this can be explicable by the use of wood fire to heat the room.

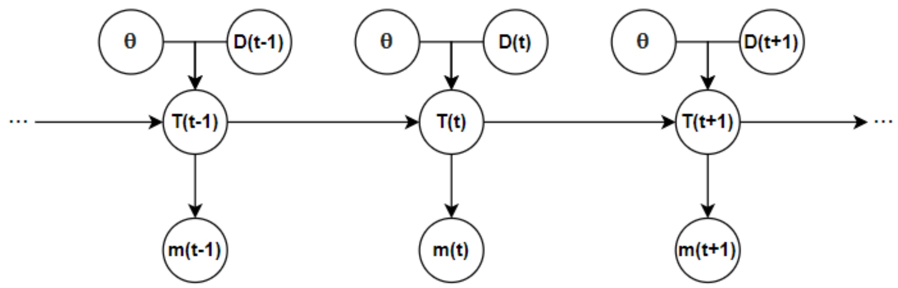
We could therefore also have tried to take into account the fact that fires were conducted during cold periods.

About the thermal influence between rooms we have look at the correlation between all the room's temperature of the house in order to make a connection between the more correlated ones. However results where not really concluding and biased due to the influence of outdoor temperature and some internal heats. We therefore decided to try all combination and choose the best one as explain later.

3 Description of the model

3.1 Model as a Markov Chain

Our model can be represented as a `Markov Chain`:



Where, at time t:

- $\mathbf{T(t)}$ is the predicted temperature from our model
- $\mathbf{m(t)}$ is the measured temperature from the sensor
- θ is the set of parameters of the model
- $\mathbf{D(t)} = [T_a(t-1), T_w(t-1), T_s(t-1), T_{neighbors}(t-1), T_r(t-1)]$ is the data available

3.2 Transition model

The transition from one time-step to another is supposed deterministic and is determined by the following equation:

$$T_z(t+1) = \left(\frac{T_a(t) - T_z(t)}{R_z \times C_z} + \frac{Q_z(t)}{C_z} - \sum_{n \in neighbours} \frac{T_z(t) - T_n(t)}{R_{z,n} \times C_z} \right) \times \Delta t + T_z(t) + a_z \quad (15)$$

Where $Q_z(t) = cp \times m_z \times (T_r(t) - T_z(t))$ and

$$T_r(t) = \begin{cases} T_w(t), & \text{if } T_{s,z}(t) - T_z(t) > 0 \textit{ (radiator on).} \\ (T_r(t-1) \times k_z), & \text{otherwise } \textit{(radiator off).} \end{cases} \quad (16)$$

with $(0 < k_z < 1)$.

About the **temperature set** and the **water temperature** we consider them, such that when the temperature set is higher than the temperature of the zone we simply **turn on the radiator** (temperature of radiator is estimated by the temperature of the water), in the other case we **turn it off** but taking into account the thermal inertia[1] that contributes to additional heating of the room with a parameters k s.t. $(0 < k < 1)$.

Also concerning equation *2*, we have first tried to add a parameters *tolerance* such that: $-1 \leq tolerance \leq 1$ instead of using 0 but when optimizing our model its value was at the order of 1^{-20} therefore for simplicity and efficiency reason we preferred let the constant 0.

[1]We had been advice to change the formulation of thermal inertia by $T_r(t-1) - k_z \times Q_z(t-1)$, however it does not seems to work well in our case therefore we decided to keep what we had and try that formulations again for milestone 4 with a new dataset

3.3 Temperature measurements

A stochastic relationship between measurement m_{t+1} and model parameters θ can be expressed as:

$$m_{t+1} = T_z(t+1)(\theta) + \epsilon$$

Where ϵ represent the noise that accounts for measurement error, and the distribution of the noise is assumed to follow a Student's t-distribution (see later in the report).

The thermal model as a Markov Chain tells us that the measurement

$$m_{t+1} \sim P(m_{t+1} \mid \hat{T}(t+1)(\theta))$$

can be written as

$$m_{t+1} \sim t(m_{t+1} \mid \hat{T}(t+1)(\theta), df)$$

Where $\hat{T}(t+1)(\theta)$ is the predicted temperature of our deterministic model.

4 Fitting procedures

In this section, we are going to present to you the best configurations / assumptions that worked the best for our model.

4.1 Assumptions about the rooms

Below is how the house is organized, and how the rooms are segmented.

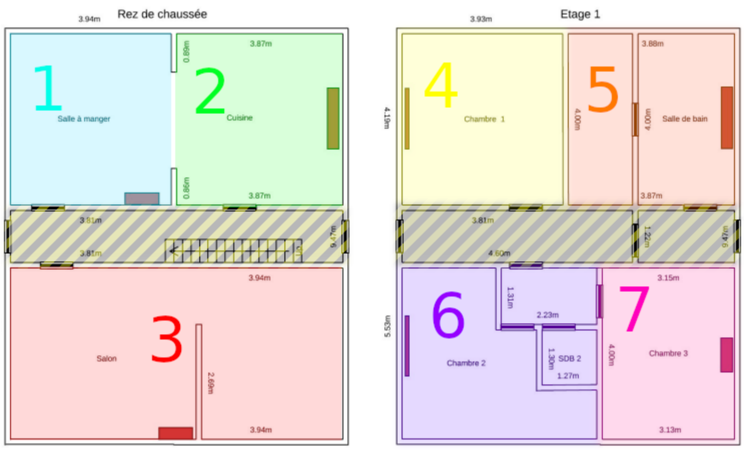


Figure 1. Room segmentations

Some assumptions are made:

- Walls that aren't delimiting the area are not considered. For instance, we don't take into account the wall splitting the area 5 into two parts.

- Halls are not considered. Since we don't have any data about the temperature in the hall, we make the assumption that it shouldn't disturb the temperature of the areas.
- Zones are considered to be completely closed. Indeed, for instance, hole between the area 1 and the area 2 would not be taken into account. For our model, this hole will be represented as a wall for which the resistance would probably end up to be very low. (high interaction between those two areas)

4.2 Interactions between rooms

The rooms of the house are grouped into three clusters. These clusters have been chosen such that this configuration ended up to be the best configuration for which the total `RMSE` is the lowest.

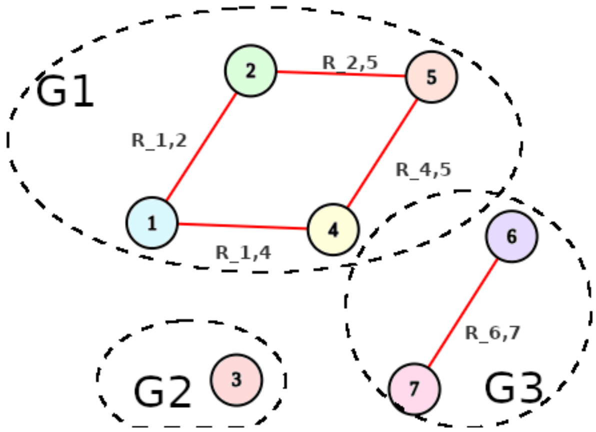


Figure 2. Rooms organized in clusters

Therefore, this is supposed that a room r_1 thermally influences another room r_2 if r_1 and r_2 are contained in the same cluster. There is no thermal influence between two clusters.

This allow to optimize parameters in cluster mutually exclusively in order to speed up the computation and have result less biased.

4.3 Parameters estimation

We seek to minimize the following objective function that is the `negative log likelihood of the Student's t-distribution`. That is, for a given cluster, the following minimization problem:

$$\begin{aligned} \min_{\theta} \quad & \sum_{r \in Cluster} \left(\sum_{t=1}^{N} -f(T_{rz}(t), \hat{T}_{rz}(t)(\theta), 1, df) \right) \\ \text{s.t.} \quad & 0 < k < 1 \end{aligned} \tag{17}$$

Where $f(x, \mu, \sigma, v)$ is the log-probability of drawing the sample x given mean μ, scale σ and degrees of freedom v. of mean μ, scale σ, degrees of freedom v and evaluated for a value x. θ represents all the parameters associated to the cluster and df. Regarding the scale, we kept it equal to 1 because the parameter *df* already controls the thickness of the tails.

For parameters estimation we use observed value at time t in order to predict the temperatures at time $t + 1$.

4.4 Distribution of the residuals

The way we estimate our parameters highly depends on our guess about which distribution does the noise follows in the measurements. In the optimization process, we assume that the measurement of the sensors follows a `Student's t-distribution`, where the degrees of freedom *df* is estimated.

At the beginning, our first guess was to make the assumption that the noise was following a `Normal distribution`. However, when looking at the histograms of the residuals, they were not following a normal distribution at all.

After doing some researches, the best distribution that was explaining our data the **best** was a `Student's t-distribution`, where the degree of freedom *df* has been estimated to a value close to 10. The fact that *df* is not a high value reinforce the fact that assuming a `Gaussian distribution` was a bad guess.

Below is presented the histograms of the residuals for all the rooms. Except the living room, the other rooms' residuals follow the guessed distribution. Indeed, the living room is a bit special because there are more outliers for which our model can not explain, because lack of information (such as the wood fire, making the zone temperature higher for no reasons, from the model's perspective).

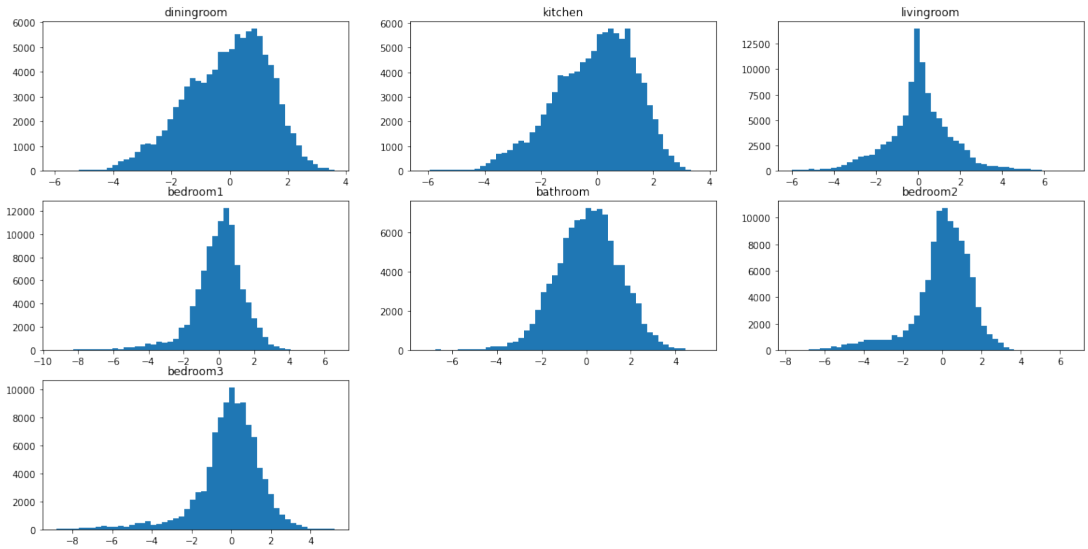


Figure 3. Residual of our predictions given observed temperatures

4.5 Results

In this section we will look at the results when doing prediction for 1 year and for a typical winter week.

- **One year**:

  Showing the result of a 1 year prediction is not the most important for that project as at the end the idea is to predict the temperature for the next hours/days to set the temperature set. However, it is still interesting in a sense to see if our model arrive to describe the thermal physics of the house

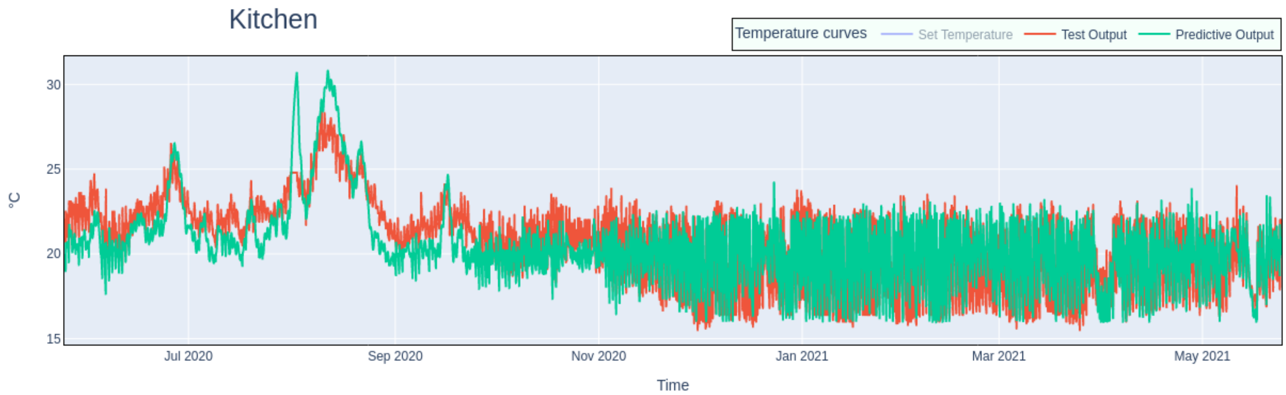


Figure 4. RMSE between observed/predicted temperatures: 1.44°C

  In average for each room we have a RMSE of 1.49°C.

- **Typical winter week**: By looking at this plot, we can see that our model seems to follow the trend even if it is not perfect, mainly because we have made the assumptions of a linear thermal loss of the radiator's heat.

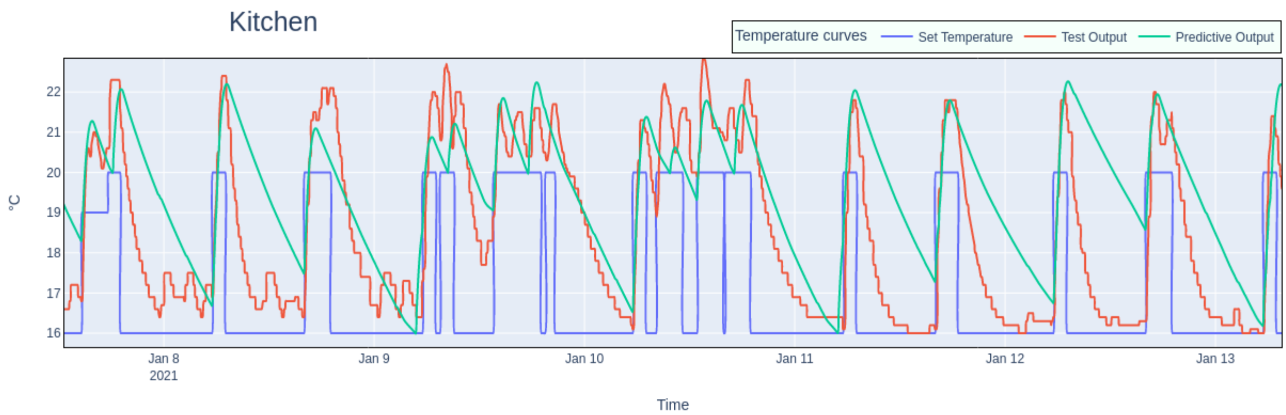


Figure 5. RMSE between observed/predicted temperatures: 1.62°C

4.6 Uncertainty

4.6.1 Bayesian modeling

We wish to make our model probabilistic, in other words, we want to bring uncertainty on our parameters. They are not deterministic anymore, each parameter now represents a **random variable**.

In addition to that, we want to see how confident our probabilistic model is regarding the parameters and given the data. Formally, we wish to compute the `posterior` probability

$$P(\theta|m_{t+1}) = \frac{P(m_{t+1} \mid \hat{T}(t+1)(\theta)) \times P(\theta)}{P(m_{t+1})} \tag{18}$$

Where `likelihood`3.3 is known. The `prior` must be defined by us, as we are not expert in the field the prior probability of each parameter $p \in \theta$ is uniformly distributed in an interval for which seems us to be the more convenient by looking at the literature[2].

However, the `evidence` is too complex to compute. Fortunately, we can sample from the conditional posterior distributions and, after convergence, actually draw samples from the joint posterior distribution

4.6.2 MCMC

We will use the `Monte Carlo Markov Chain` method in order to estimate out posterior distribution.

To capture uncertainty, one subset of the dataset (basically, the time-series of two weeks) is used. However, the model behaving very differently during the winter and during the summer, they will be evaluated separately[3]. Indeed, during the summer, the radiator is not used.

For each cluster, MCMC is ran with a number of walkers being twice the number of parameters of the cluster, each walker starting at different positions that are very close to $\theta*$ (where $\theta*$ are the optimized parameters) and making 5000 steps, where the first 1000 steps are discarded.

The idea is to have a sliding window of two weeks, that will be used in order to predict the next few days and capture uncertainty on these predictions. For evaluation purposes, we are still using the real values of T_a, T_z and T_w as inputs for predictions.

The starting point is the last temperature measured and we plot with a confidence interval of 95%.

We can see in figure **6** that the temperature measurement are almost always comprise in the 95% confidence interval of the posterior which is a good news. We unfortunately do not have enough space to talk about the results of the corner plot but it is really close to what we have found in the last milestone.

[2]Sizing and Operation of an Isolated Microgrid with Building Thermal Dynamics and Cold Storage - Selmane Dakir, Ioannis Boukas, Vincent Lemort and Bertrand Cornélusse

[3]See previous milestone for results

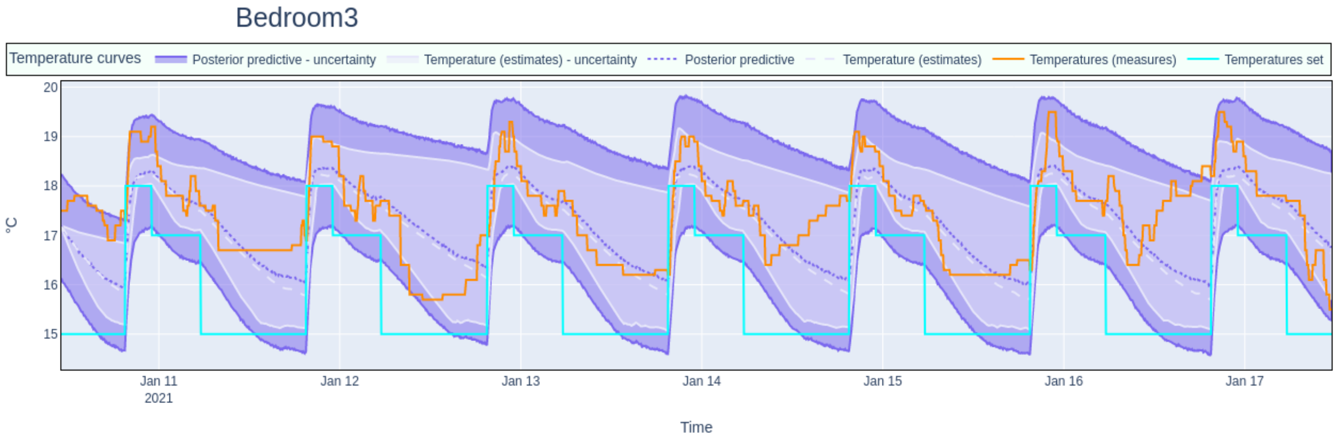


Figure 6. Prediction of a week in January given the two precedent week, for the Bedroom3

5 Parameters value

Room (z)	C_z	R_z	R_xy	m	a	k
Dining room	148500	0.8	{0.1, 30}	0.001	-0.005	0.92
Kitchen	244800	0.5	{0.1, 0.54}	0.002	0.009	0.93
Living room	8700	30	$\{\emptyset\}$	$1.75 * 10^{-05}$	0.004	0.9
Bedroom 1	591300	0.4	{30, 0.52}	0.002	0.006	0.9
Bathroom	33900	3	{0.54, 0.52}	0.0001	0.0009	0.91
Bedroom 2	350400	0.7	{0.2}	0.0004	0.002	0.9
Bedroom 3	416400	1	{0.2}	0.0005	0.007	0.9

Table 1. Estimated value of unknown parameters, R_{xy} values are sort according to **Figures 2**

By looking at literature our values of C seems under evaluated and the one of R over evaluated, this can be explainable by the fact that they are highly related in our model.

6 Conclusion

In conclusion we are quite proud of what we have done and it is fascinating to see that with just a data set and some knowledge about physics, we can end up with a model that can explain the whole data and be close to *reality*[4], however our model is still subject of few improvement such as including internal heating, solar irradiance and updating the thermal inertia formula.

But in the end we have learned a lot of things with this project even if we had never heard about heat transfer and thermal physics before. And we really look forward milestone 4[5] to use what we have done for a more practical case!

[4]Reality is stochastic
[5]If we survive from exams