A hybrid simulation/real data based learning approach for building thermal comfort modeling

Sara Yasmine Ouerk, Romain Barbedienne, Benoit Charrier, Hassan Bouia, Mouadh Yagoubi, Thierry Duforestel

A hybrid simulation/real data based learning approach for building thermal comfort modeling

Sara Yasmine Ouerk [1], Romain Barbedienne [1], Benoit Charrier[2], Hassan Bouia[2], Mouadh Yagoubi[1], Thierry Duforestel[2]
[1] IRT SystemX, Palaiseau, France
[2] EDF R&D -EDF Lab les Renardières, Moret-Loing-et-Orbanne, France

Abstract

Machine learning (ML) methods have recently received a growing interest in thermal comfort modelling thanks to their ability to integrate the human factor through real data. However, the main drawback of these methods remains the large amount of high-quality expensive data required to train the learning model. These data require a large number of attendees, with specific thermal conditions. In this paper, we propose an approach for reducing the cost of data acquisition by hybridizing real survey data with simulated thermal data. We validate the proposed approach by benchmarking several ML models with different configurations to predict two classes namely: comfort and discomfort. Obtained results have shown that the random forest model outperforms other ML methods and achieves promising results even on the discomfort which is generally less represented by the data for thermal comfort prediction.

Highlights

- Dwelling thermal comfort prediction
- ML/Simulation hybridization for data augmentation
- ML/Simulation hybridization for feature enrichment
- Benchmarinkg and model selection

Introduction

According to the International Energy Agency (IEA) (IEA, 2022), the building sector is the first responsible of energy consumption (30%) and greenhouse gas emissions (27%) in the world, and thermal performance has become a critical consideration in the design of new buildings. In the last few years, improving the performance of the building stock, particularly through the renovation of existing buildings, has received a particular attention. As the energy performance of buildings increases, occupant behavior is becoming an increasingly important factor.

The behavior of the occupants has a direct impact on energy use and therefore on the total energy demand of the building. For example, in the residential sector, the energy consumption of two dwellings with the same intrinsic characteristics (e.g. geometry, energy performance of buildings and energy systems, other electrical equipment,…) can be quite different depending on the composition of the households. Habits and activities can vary greatly from one occupant to another and are influenced by the characteristics of the household to which they belong (e.g. single person, couple with children, retired couple).

In order to make energy performance predictions more reliable after renovation and limit the rebound effect, it is therefore important to finely model occupancy and produce personalized analyses according to the profile of the occupants.

Among the influential factors linked to occupancy, the concept of occupant's thermal comfort has a first-order impact on heating consumption, as it has an important impact on the actions of occupants controlling the energy systems, and thus on the building energy consumptions. The ability to realistically model thermal comfort is a key to assess correctly the energy consumption of a housing.

In this work, we describe a new hybrid methodology for predicting perceived thermal comfort using machine learning models based on both survey data and physical models.

Related work

Thermal comfort prediction

The international organization ASHRAE defines thermal comfort as the "condition of mind that expresses satisfaction with the thermal environment and is assessed by subjective evaluation" (Standard, 1992). The prediction of thermal comfort is a great challenge for dimensioning thermal systems or for estimating the consumption of households.

Human perception of thermal comfort is usually predicted using Predicted Mean Vote model. This model also known as the Fanger model (Ekici, 2013) that has been established experimentally (Fanger & others, 1970),is widely used in industries, and was standardized in EN ISO 7730 (Standardization, 2005).. However, many limitations have been raised since its establishment. The first one is that Fanger model is not subjective enough for individual comfort prediction (Schaudienst & Vogdt, 2017) (van Hoof, 2008). Another weakness of this model, is that Fanger model has been defined in Laboratory conditions. However, Oseland (Oseland, 1995) has shown by experimental study that the comfort felt at home or in the office does not correlate with the comfort felt in a climate chamber.

The second class of model used in order to predict comfort is the physiological models. These models calculate the skin temperature. This calculation is performed using in one hand the laws of thermodynamics of the body named

the passive part and in other hand the regulation behavior of body named the active part. The most important challenge of such models is to simulate the whole human regulatory behavior (Hensel & Schafer, 1984), and the Stolwijk model (Stolwijk, 1971) was one of the most popular models proposed to this purpose. The passive part is composed of 25 nodes. Each segment between nodes represents a body part, for example a leg. In contrast to the Fanger model, Stolwijk model allows to consider the transient comfort. But this model is still not very subjective, because it would be necessary to measure precisely each segment and the variables of the active part.

The recent advances in machine learning have made it possible to create more subjective models of thermal comfort (Sundaram et al., 2021). Farad et al. (Fard et al., 2022) has recently shown that Machine learning models could outperform the PMV model by up to 35.9%. They also raised the limitations of machine learning models for comfort modeling. The main limitation of such a model is the quality and the large amount of data required to obtain good predictions. However, thermal comfort data is expensive to obtain because it requires a lot of attendees, facing sufficiently representative thermal conditions. The international organization ASHRAE has made available a database of thermal comfort. This database contains the thermal comfort information of many people in a large number of countries. Although this initiative is very promising and allows to improve the predictive models, it is reported that a significant amount of information is missing. Moreover, this set of data is global, and requires a transfer of knowledge to target the prediction of thermal comfort for a specific population. Gao et al. (Gao et al., 2021) used a transfer learning approach using ASHRAE dataset to predict thermal comfort of inhabitants of specific USA city dataset. The accuracy is only 3% higher with transfer learning than with a predictive model based on random forest model.

Data augmentation

One solution to increase the quantity of data explored in this paper is data augmentation.. , and we will focus on data augmentation techniques applied to time series. We will assume that the thermal comfort data are a time series. For each time step, a comfort label is given.

Wen et al. (Q. Wen et al., 2020) propose a survey on data augmentation based on time series. They introduce a taxonomy of time series data augmentation techniques. These approaches consist in applying a simple transformation of data. For example, reverse the time to create a new time series, or even crop the time series to generate shorter series. These approaches are not adapted for thermal comfort, because it could increase noise, or even generate wrong data.

The most promising approaches for the thermal comfort domain, are advanced approach, especially Machine Learning technics. Quintana et al. (Quintana et al., 2020) apply a Generative Adversarial Networks (GAN) to address the problem of unbalanced data in comfort modeling. GAN improved the prediction of machine learning models for the tested datasets between 2 in 17%.

Hybridization of ML model and simulation

Simulation is adapted in order to increase the quantity of data. For example, the training of autonomous car requires an important dataset. Dosovitskiy et al. (Dosovitskiy et al., 2017) proposes a car simulator in order to enrich data in the context of autonomous driving. Another example for data augmentation with simulation is in system biology context (Deist et al., 2019) . The real data are mixed in order to create new samples. These samples are then simulated, and a clustering algorithm is trained on enriched data.

Chinesta et al. (Chinesta et al., 2020) introduced the concept of hybrid twins. This model addresses the lack of data from a different perspective. The model completes the results of physical model with a corrected term. This corrected term is learned by a machine learning algorithm. In this context, machine learning models can be simple and require less data than a complete model.

These hybridizations techniques are adapted for a physical phenomenon. However, in the context of thermal comfort, there is no suitable physical model, because thermal comfort is a human feeling.

Hybridization of ML model and simulation

The proposed approach is to hybrid simulation model with Machine learning model. The simulation model generates context variables that describe the environment. Applied to the field of thermal comfort, the simulation model will calculate the thermal environment variables. Thus, for each time step, context variables are generated. But the main objective is to increase the amount of data. The generated time steps are not in the original dataset. Thus, the variable to predict must be evaluated for these generated time step. For this purpose, a second simulation model or an estimator is needed to evaluate this variable.

The description of the proposed approach and a comparison with classical approach is described in Figure 1.

X_{DB}, are the variables used to generate the simulation environment, to simulate target variable and as an input for the learning mode, y_{DB} the target variable in the original dataset. Let f_{ENV} be the application which simulate the environment; $f_{ENV}: X_{DB} \longmapsto X_{PHY}$, where X_{PHY} are the variables generated by simulations. Let f_{PHY} the second simulation model; $f_{PHY}: y_{DB}, X_{DB}, X_{PHY} \longmapsto y_{PHY}$ where y_{PHY} is the target variable. Finally, let f_{LEARN} to be he machine learning model; $f_{LEARN}: X_{DB}, X_{PHY} \longmapsto \hat{y}$, f_{LEARN} is learned using y_{PHY}.

The generation of context variables has three objectives. The first one is to increase the amount of data, as a time-based simulation model can calculate features at different time steps. The second one consists in improving the prediction of the learned model by adding non-available features. Finally, the third objective is to reduce the gap between the training dataset and the one that will be used for the inference and evaluation. In fact, the classical comfort models are

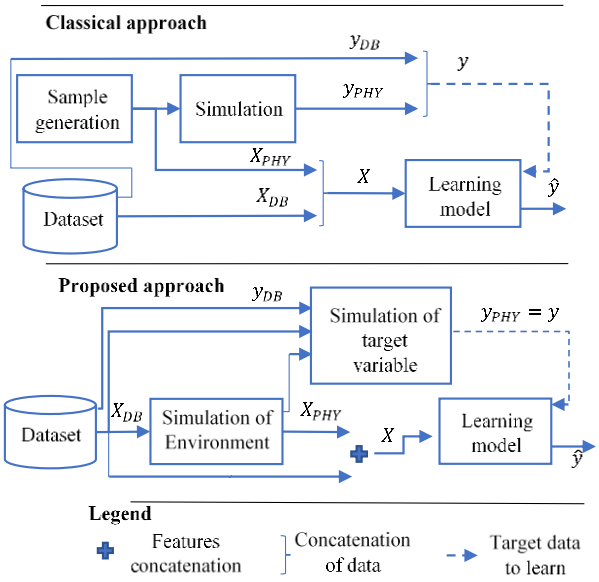

Figure 1 Comparison of classical usage of simulation and the proposed method

learned from real data, but the evaluations of these laws are done by simulations. The results of simulations can differ from the real models. Thus, a learning process based on simulated data will be more accurate if the inference of the learned model is made with simulated variables.

However, the simulations are required to be accurate to avoid biasing the ML model. The difficulty is to build a simulation model that is as close as possible to the real context. We will describe this part briefly in the application section.

Application

The process of data enrichment by simulations will be illustrated with the prediction of the thermal comfort of the inhabitants of a dwelling. This study will be limited to the prediction of the sensation of cold during the heating period (October 1st to April 30th).

Data

Data description

In this context, a survey was conducted among 4,000 French households. The aim of this survey was to link the comfort data to the characteristics of the dwellings, the heating systems, the heating habits of the occupants as well as to some social and economic characteristics of the household (number of inhabitants, their ages, genders, professional activities, incomes ... etc.). Inhabitants also indicated their comfort levels, with 5 possible answers: the person is never cold, is cold for at least 24 hours, is cold for a few days, is cold almost all the time or is cold all the time. Statistical analysis of the questionnaire showed that 85% of the inhabitants are comfortable all the time, which creates a notable imbalance between the comfort classes.

These data have been pre-processed, notably by deleting dwelling with inconsistent or missing values, making restrictions on the types of dwellings and respecting the minimum and maximum number of rooms (due to simulation modeling, see the paragraph on simulation). After several preprocessing, we ended up with a dataset of 1394 dwellings.

Limitations of the survey data for comfort learning

Due to the small amount of data, EXtreme Gradient Boosting (XGBoost) model was trained on the survey data to predict the different comfort classes of the dwellings. The characteristics used for learning are related to the type of heating, the characteristics of the dwelling and the inhabitants (ex: surface, age or gender of inhabitants, …etc.), and the setpoint temperatures (mean and standard deviation over a typical week). Table 1 shows the results of the classification on the test set. The model only manages to recognize the majority class "Never cold", which has the highest support. For the other classes, the performance is null, except for the class "Few days" which has a slightly higher support than the rest of the classes, which indicates that the amount of data is not sufficient to learn the specificity of these classes, thus requiring an increase in the amount of data.

Table 1 XGBoost classification results on survey data

Class	Metric	Precision	Recall	F1-score	support
Never cold		0.89	0.98	0.94	246
24h at least		0.00	0.00	0.00	4
Few days		0.50	0.07	0.12	15
Almost all the time		0.00	0.00	0.00	11
All the time		0.00	0.00	0.00	3

Simulation

As described in the previous section, the simulation models require to be accurate to avoid biasing the ML model. The general methodology requires two simulations.

Environment simulation model

The first simulation model consists in generating environments variables. In the context of Thermal comfort, the environments model should generate air temperature variables, air velocity variables, relative humidity variables and mean radiant Temperature variables. For this first version, which allows to validate the feasibility of the concept, we realized thermal models of housing allowing to calculate the variables of air temperature, radiated temperature, radiative heat flow and convective heat flow.

For this model each dwelling in the survey will be simulated. These thermal models are 1D simulation models. Simulations of dwellings are realized with BuildSyspro (Plessis et al., 2014) which is a Modelica library developed by EDF. Each room of the dwelling is considered as a thermal node. We will assume the temperature of each room to be homogenous.

The simulation models are generated using two templates., The first one is an apartment composed of a living room, one or two bedrooms, a bathroom, a kitchen and a toilet. This template is named “Matisse” Apartment. The second one is a House composed of a living room, two or three bedrooms, a bathroom, a kitchen and a toilet. This template is named “Mozart” House. These two templates represent 37% of the data set. The other housing types will not be used for learning.

The information used to fill these simulation templates are: for the dwelling; House insulation, year of construction, and for each room; the heating system parameters (energy, power, technology used), the surface, the orientation and the surface of the windows. These parameters are described in the dataset.

The inputs of the model are generated with two kind of information described in the dataset. The first one is the heating habit, including the setpoint temperature of the heating systems for each room, the times when the windows are open and the times when the shutter are closed. All these data are given in the survey hour by hour over a week for each room. The second one is the meteorological condition variables given by RT2012 which is a French standard of thermal regulation. These variables include air temperature, direct and diffuse radiation, sky temperature, wind speed and sun position. These variables are adjusted every 30 minutes and for a period of one year. A set of data is calculated for 8 different French climates. The set of data chosen depends on the location of the dwelling which is filled in the survey. The complete simulation workflow is descried in Figure 2.

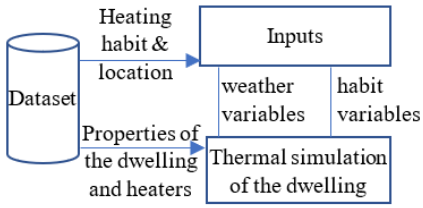

Figure 2 Description of simulation workflow

An algorithm was realized in order to fill the simulation template and generate input information for each dwelling of the dataset. Output variables include Operative temperature, radiation temperature, air temperature, radiant heat flow and convective heat flow which are calculated for each room.

Simulation of comfort model

The second simulation model simulates the comfort variables. The main issue considering comfort modeling is that no simulation model can give an exact solution. The most rigorous approach would be to use a physiological model, such as the Stolwijk model. The Stolwijk model could be adapted for each inhabitant of the dwelling using the sociological description of the dataset (gender, age, …etc.). The environment simulation model could feed the Stolwijk model in order to estimate the skin temperature, and evaluate if the residents are comfortable or not.

For this first version, the time steps for which the inhabitants are uncomfortable are approximated. The dataset contains information on the duration of discomfort of the inhabitants. The inhabitants filled if they are cold for more than 24 hours, for a few days, almost all the time or all the time. Resident comfort survey data is collected based on a 24-hour period. This means that a resident who is cold for a short period of time will be considered comfortable all the time. The learning model will not take transient effects into account. The proposal was to calculate for each logging a threshold, based on operative temperature simulation. If operative temperature is below this threshold, inhabitants are considered as uncomfortable, if operative temperature is above the threshold, inhabitants are considered as comfortable. The threshold is calculated in order to satisfy the discomfort duration set in survey. Note that this simple model is a proof of concept. The objective is to ensure the feasibility before proposing an improved version.

Benchmarking of ML models

Time-independent modeling

Given a dataset D, of N housing, each one composed of 10128 time-steps (7 months, with a time step of 30 minutes). Each time step, is composed of dynamic features coming mainly from the physical simulation, and static features coming from the questionnaire data, representing some characteristics of the housing (like age and gender of the reference person, household income... etc.). The reference person is defined as the head of the family

The survey contains also the presence data in the different rooms of the housing for each time step. In the case where there is no presence in any of the rooms, the comfort is considered as unknown.

Figure 3 shows the class distribution over all observations. This diagram shows that that classes are unbalanced, especially for the 'discomfort' class which is very poorly represented and which makes it difficult to capture. There are also not many observations for the class 'unknown', but this class is easy to recognize since it is correlated to the presence in the rooms.

A first version was built, considering the time steps independent between them. The comfort at a given time step depends only on the simulated temperatures and the characteristics of the housing, and does not depend on the comfort at the time step that precedes it.

For this configuration, the dataset D is represented as a set of pairs $\{X_{it}, Y_{it}\}$, where X_{it}, contains the static and dynamic variables at time step t of housing i, and Y_{it} the corresponding comfort value.

To train the different models, all time steps of the different housing were randomly mixed and then divided into 3 sets (60% on Train, 20% on Validation and 20% on Test).

Several types of models were trained: Ensemble models like Random Forest and XGBoost, neural networks: Multi-Layer-Perceptron (MLP) (Singh et al., 2016), and their performance in terms of accuracy and F1 score (*Magician’s Corner*, n.d.) were compared.

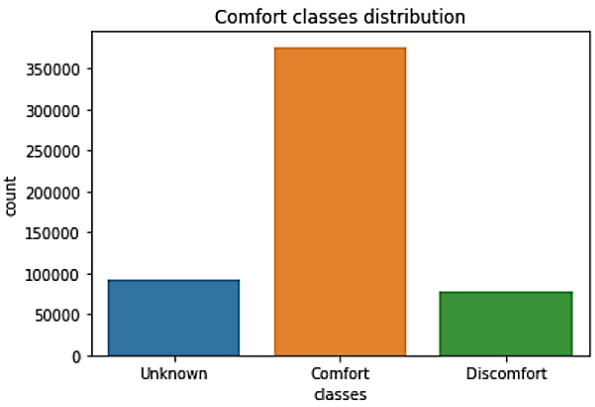

Figure 3 Comfort classes distribution

For the MLP model, some methods were tested to try to overcome the problem of class imbalance. In particular, by adding weights to the CrossEntropy loss (Wang et al., 2022) to penalize it more for the minority class. In order to balance between the 'discomfort' and 'comfort' class, the UnderSampling method by randomly removing examples from the majority class has also been tested.

Results

Table 2 shows the accuracy and F1 scores for comfort prediction with the different models listed above, for each class.

For the class 'discomfort', for which only few observations are available, the Random Forest model gives the best results, both in terms of precision and recall (thus the f1 score). The XGBoost model also gives good results, but the MLP model is outperformed for this label (-33%) for the F1-score, despite attempts to reduce the imbalance between the classes.

Table 2 Comparison of different Machine learning models for comfort prediction

Class	Model	Precision	Recall	F1-score
Comfort	MLP	0.95	0.98	0.97
	XGBoost	**0.999**	**0.999**	**0.999**
	Random Forest	**0.999**	**0.999**	**0.999**
Discomfort	MLP	0.61	0.37	0.46
	XGBoost	0.97	0.95	0.96
	Random Forest	**0.999**	**0.999**	**0.999**
Unknown	MLP	**1.0**	**1.0**	**1.0**
	XGBoost	**1.0**	**1.0**	**1.0**
	Random Forest	**1.0**	**1.0**	**1.0**

For the 'comfort' class, the Random Forest and XGBoost models both performed very well. The MLP model seems to be relatively worse but the results remain satisfying.

For the last class "unknown" the three models have very good performances, indeed, this class is easily detectable as it is correlated to the presence in the housing.

Random Forest is generally considered as the best classification algorithm for small datasets (Hu et al., 2019). It has also been shown to have the highest prediction accuracy for thermal sensation (Luo et al., 2020). In addition, it can handle high-dimensional features and judge the importance of features.

Feature importance

The importance of each variable in predicting comfort for the random forest model was measured, using the average decrease in impurity calculated from all decision trees in the forest. Figure 4 shows the 7 most important features. Among these features we find in particular the sociological criteria (income, age... etc.) as well as the operating and setpoint temperatures.

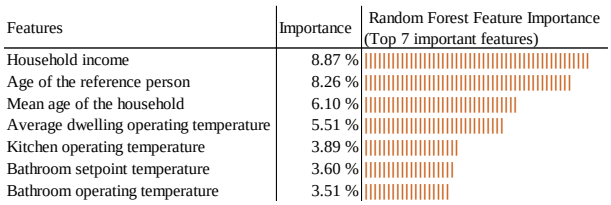

Figure 4 Random Forest feature importance

The most important feature being the household income (which should not be the case), is probably due to the fact that the more people are financially comfortable the more they will tend to better insulate their homes or increase the set temperatures without worrying about the costs involved to feel more comfortable. Then, the most influential parameters are the operating age and operative temperature. This is consistent with the results of other studies on comfort. However, we can note that the gender (male or female) does not play an important part in the features. This result is probably due to the fact that we limited the study to medium-sized dwellings generally composed of a mixed household (male and female).

Time-series modelisation

Motivation

In a second step, a second version of model prediction was built. It would allow to consider that each comfort value at a given time step depends on its previous values, in addition to exogenous variables (temperatures...etc.).

To verify this hypothesis, an autocorrelation test has been realized for the comfort series. This test is defined as the correlation of the comfort values of the sequence with the comfort values from the previous steps called lags.

Figure 5 shows autocorrelation function (ACF) for comfort values in a randomly selected household for 20 lags. The height of each spike shows the value of ACF for the corresponding lag.

The ACF, that rises above or falls below a confidence interval (the blue region in the graphic) is said to be significantly autocorrelated. In this example, all the 20 spikes are statistically significant, indicating that a comfort value at a given time step is highly correlated with the 20 values preceding it.

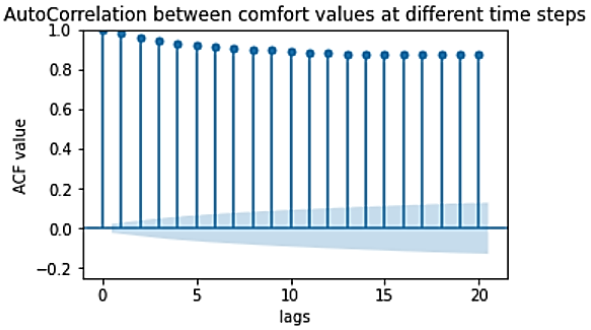

Figure 5 ACF for comfort in a randomly selected household

As the time series have sizes of 10128 time-steps, we have to split them into sub-sequences of smaller sizes using a sliding window. The window size is adjusted based on how far the comfort values are related to their past values. Thus, for the example above, based on the ACF results, we can take 21 as the window size.

In general, the values at which the spikes are no longer statistically significant were calculated for each household. In more than 96% of the cases, this value is above 20, a window size of 21 (20 time-steps in the past horizon + one prediction step) seems to be appropriate (this choice of size 20 was made to avoid having very large windows), which corresponds to a past horizon of 10 hours.

Model description

The time series are modeled using a multi-horizon model (R. Wen et al., 2018), consisting of a past horizon and a prediction horizon. This model is particularly well suited for time series prediction. It is composed of a past horizon containing in addition to the contextual information from the past, the corresponding comfort values. In the prediction horizon, only contextual information are available, and the model will predict the corresponding comfort values. Thus, the model should be able to best infer comfort in the prediction horizon (the t^{th} comfort value) from the historical data (*0, …, t-1*) and the current context, as described in Figure 6.

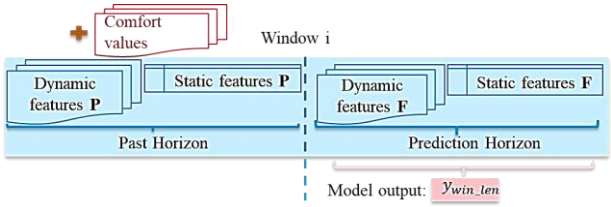

Figure 6 General scheme of the prediction model

The model is composed of fully-connected layers for the encoding of static features and GRU (gated recurrent units) layers for the encoding of dynamic features as described in Figure 7. The cost function should be calculated only over the prediction horizon.

Experimental setup and results

The multi-Horizon model was trained, with sequences of size 21, with a past horizon size of 20, and a future horizon size of one time-step. The train, evaluation and test sets have been constructed so that all subsequences of a household sequence belong only to one of these sets. The model converged after a few training epochs.

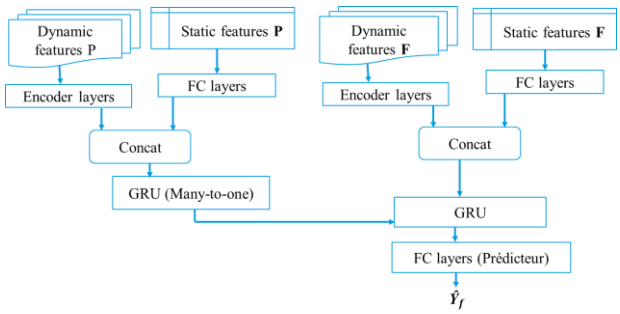

Figure 7 Multi-Horizons model architecture

To test the model, we proceeded in two different ways. First, it was fed by input sequences with the real values of comfort in the past horizon so that it predicts the last value of the subsequence. But in the context of simulation, it will not be possible to have the real comfort value at each time step. Then, in a second step, we assumed not to know the real values of comfort in the past horizon (which corresponds to our case). Thus, for each of the sequences of a household, we give it only the first 20 real comfort values, so that it predicts the 21th one, and then we give this predicted value as the last entry of the past horizon of the next subsequence and so on until going through the whole sequence, as described in Figure 8.

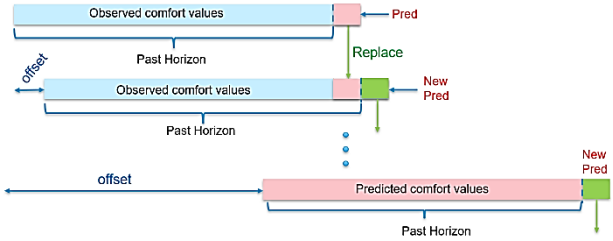

Figure 8 Iterative prediction for Multi-Horizons model

Table 3 shows the results on the test set with the two different configurations. We obtain very good performances for the different classes with the 1^{st} configuration where we suppose to know the real comfort values in the past horizon. These performances are degraded with the 2^{nd} configuration, where the comfort values in the past horizon are not known anymore. Indeed, if the model makes errors at a given time step, this will propagate throughout the sequence, which thus deteriorates the performance.

Figure 9 shows an example of comfort evolution for a person being cold for a few days, as well as the corresponding predictions with the different prediction configurations. This chronogram is defined at different time steps during the heating period. For this example, the predicted comfort labels with the configuration where the real comfort values in the past horizon are available, match perfectly the real values. On the other hand, those predicted recursively are less precise (the model predicts that the person is always comfortable).

This modeling would therefore be better suited to the case where comfort values are available per time step, and not

for this scenario where comfort should be predicted for the whole heating period without having any information on comfort at previous time steps.

Table 3 Multi-horizons model classification results

Class	Prediction strategy	Precision	Recall	F1-score	Support
Discomfor	Real values in past horizon	**0.999**	**0.999**	**0.999**	33550
	Recursive prediction	0.88	0.25	0.39	
Comfort	Real values in past horizon	**0.999**	**0.999**	**0.999**	124390
	Recursive prediction	0.83	0.99	0.90	
Unknown	Real values in past horizon	**1.0**	**1.0**	**1.0**	44220
	Recursive prediction	1.0	1.0	1.0	

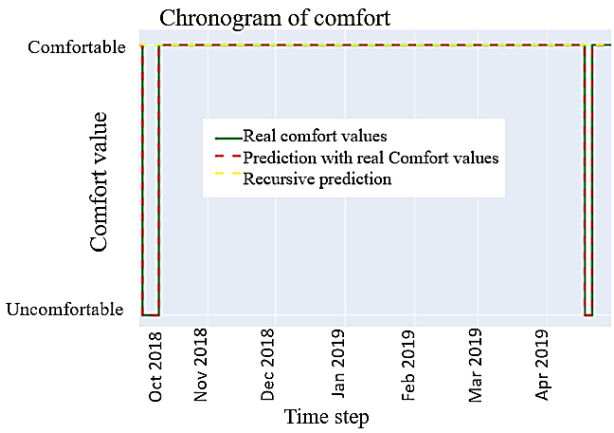

Figure 9 Chronogram of comfort

Discussion

The above results show that comfort prediction using a random forest model gives very promising results.

However, this first version is based only on the temperature, convective and radiative flux variables. The variables of humidity and air speed have been neglected. Moreover, the comfort is described as a threshold related to the operating temperature (this threshold is calculated according to the answers to the survey). To complexify our approach, a more realistic comfort model based on a hysteresis cycle has been tested. This model allows to consider the comfort inertia on some time steps. The results of the random forest model were similar to those obtained with the first version. On the other side, as described in the paragraph about the simulation of the comfort model, a Stolwijk model could be the next step to test. However, this model requires the simulation of humidity, air velocity, the activity and clothing of each person in the house. A coupling with a thermal model including the calculation of humidity and air velocity could improve the reliability of the tested model. Moreover, a coupling with a multi-agent model like the SMACH model developed by EDF would allow a more accurate prediction of comfort.

Regarding the results of the multi-horizons model, the poor results observed when using the prediction for the past horizon can be explained by the low number of dwellings used. In fact, due to the low computational resources, we were able to train this model on about fifty dwellings. Larger calculation resources will allow us to train the model on more data. However, for this model, we could observe that good prediction results were obtained using the actual comfort value at previous time steps. This model could be applied in the context of a smart thermostat that learns the temperature set point according to the comfort indicated by the user; and therefore, predict very accurately the comfort value at the next time step. This would help to change the temperature setpoint for the next time steps until the multi-horizons model predicts a user comfort. This could allow to optimize the compromise between setpoint temperature and the user comfort.

Conclusion

In this study, a hybridization between thermal simulation and data-based thermal comfort modeling was implemented. This hybridization aimed to solve the problem of lack of data and to enrich our database with variables not filled in the survey. The thermal simulations were used to generate the contextual variables of a dwelling. Several machine learning models were trained and tested on this hybridization data to compare their performance. Two modeling hypotheses were studied in order to select the one best suited to our use case. For the first modeling, the comfort values were assumed to be independent. For the second modeling, the comfort values of a time step were assumed to be dependent of previous time steps. Very promising results were attained with the first assumption, that of time step independence, using a random forest model.

The prediction of thermal comfort in this context of hybridization is a first step. The main objective is to predict the energy consumption of a building. The consumption depends on the temperature set point. Comfort is a factor, but also the household income, the price of energy, the type of housing programmer, the presence or absence of a person in the housing, or the ecological awareness of the household inhabitants. We plan for this second study to couple three models; a thermal simulation model, a multi-agent simulation model, and a machine learning model to predict the temperature set point.

Finally, we also aim to have a general vision of the different scenarios (comfort prediction as well as setpoint temperature prediction) and to evaluate the different models used to solve these problems. In this context, different generic evaluation criteria (machine learning, industrial readiness, physics, …etc.) are being implemented and will be compared with the help of a benchmarking platform including other use cases that hybridize simulation and machine learning. The LIPS

platform will be used to perform such a benchmark (Leyli-abadi et al., 2022).

References

Chinesta, F., Cueto, E., Abisset-Chavanne, E., Duval, J. L., & Khaldi, F. E. (2020). Virtual, digital and hybrid twins: A new paradigm in data-based engineering and engineered data. *Archives of Computational Methods in Engineering*, *27*, 105–134.

Deist, T. M., Patti, A., Wang, Z., Krane, D., Sorenson, T., & Craft, D. (2019). Simulation-assisted machine learning. *Bioinformatics*, *35*(20), 4072–4080.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. *Conference on Robot Learning*, 1–16.

Ekici, C. (2013). A review of thermal comfort and method of using Fanger's PMV equation. *5th International Symposium on Measurement, Analysis and Modelling of Human Functions, ISHF 2013*, 61–64.

Fanger, P. O. & others. (1970). Thermal comfort. Analysis and applications in environmental engineering. *Thermal Comfort. Analysis and Applications in Environmental Engineering.*

Fard, Z. Q., Zomorodian, Z. S., & Korsavi, S. S. (2022). Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. *Energy and Buildings*, *256*, 111771.

Gao, N., Shao, W., Rahaman, M. S., Zhai, J., David, K., & Salim, F. D. (2021). Transfer learning for thermal comfort prediction in multiple cities. *Building and Environment*, *195*, 107725.

Hensel, H., & Schafer, K. (1984). Thermoreception and Temperature Regulation in Man. In E. F. J. Ring & B. Phillips (Eds.), *Recent Advances in Medical Thermology* (pp. 51–64). Springer New York.

Hu, W., Luo, Y., Lu, Z., & Wen, Y. (2019). Heterogeneous Transfer Learning for Thermal Comfort Modeling. *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 61–70.

IEA. (2022). *Buildings*. IEA. https://www.iea.org/reports/buildings

Leyli-abadi, M., Marot, A., Picault, J., Danan, D., Yagoubi, M., Donnot, B., Attoui, S.-E., Dimitrov, P., Farjallah, A., & Etienam, C. (2022). LIPS-Learning Industrial Physical Simulation benchmark suite. *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Luo, M., Xie, J., Yan, Y., Ke, Z., Yu, P., Wang, Z., & Zhang, J. (2020). Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II. *Energy and Buildings*, *210*, 109776.

Magician's Corner: 9. Performance Metrics for Machine Learning Models. (n.d.).

Oseland, N. A. (1995). Predicted and reported thermal sensation in climate chambers, offices and homes. *Energy and Buildings*, *23*(2), 105–115.

Plessis, G., Kaemmerlen, A., & Lindsay, A. (2014). *BuildSysPro: A Modelica library for modelling buildings and energy systems*. 1161–1169.

Quintana, M., Schiavon, S., Tham, K. W., & Miller, C. (2020). Balancing thermal comfort datasets: We GAN, but should we? *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 120–129.

Schaudienst, F., & Vogdt, F. U. (2017). Fanger's model of thermal comfort: A model suitable just for men? *Energy Procedia*, *132*, 129–134.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315.

Standard, A. (1992). 55, Thermal environmental conditions for human occupancy. *American Society of Heating, Refrigerating and Air Conditioning Engineers*, *145*.

Standardization, I. O. for. (2005). *ISO 7730 2005-11-15 Ergonomics of the Thermal Environment: Analytical Determination and Interpretation of Thermal Comfort Using Calculation of the PMV and PPD Indices and Local Thermal Comfort Criteria*. ISO.

Stolwijk, J. A. (1971). *A mathematical model of physiological temperature regulation in man*. NASA.

Sundaram, M. K., Ali, N., & E, R. (2021). Building thermal simulation-based climate classification of India. *Proceedings of Building Simulation 2021: 17th Conference of IBPSA*, *17*, 805–811.

van Hoof, J. (2008). Forty years of Fanger's model of thermal comfort: Comfort for all? *Indoor Air*, *18*(3), 182—201.

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, *9*(2), 187–212.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2020). Time series data augmentation for deep learning: A survey. *ArXiv Preprint ArXiv:2002.12478*.

Wen, R., Torkkola, K., Narayanaswamy, B., & Madeka, D. (2018). *A Multi-Horizon Quantile Recurrent Forecaster*