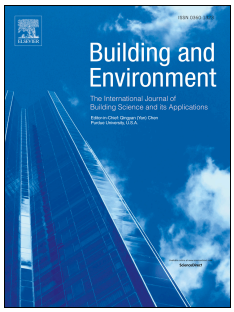
Accepted Manuscript



Reconstruction of the indoor temperature dataset of a house using data driven models for performance evaluation

Luis M. Candanedo, Véronique Feldheim, Dominique Deramaix

PII: S0360-1323(18)30247-6

DOI: 10.1016/j.buildenv.2018.04.035

Reference: BAE 5431

To appear in: *Building and Environment*

Received Date: 11 January 2018

Revised Date: 16 April 2018

Accepted Date: 26 April 2018

Please cite this article as: Candanedo LM, Feldheim Vé, Deramaix D, Reconstruction of the indoor temperature dataset of a house using data driven models for performance evaluation, *Building and Environment* (2018), doi: 10.1016/j.buildenv.2018.04.035.

Reconstruction of the indoor temperature dataset of a house using data driven models for performance evaluation

Luis M. Candanedo, Véronique Feldheim, Dominique Deramaix

Luismiguel.candanedoibarra@umons.ac.be
Tel +32(0)65 37 4471 Fax +32(0)65 37 4400
Thermal Engineering and Combustion Laboratory, Rue de l' Epargne 56, 7000 Mons.
University of Mons, Belgium

Abstract:

Whenever the long term monitoring of a building is attempted it is likely that specific sensors or the whole monitoring system used may experience long-term failure therefore creating important gaps in one or more variables of special interest. These long gaps may not be addressed using simple linear interpolation. The option of only using the available data for descriptive statistics would produce results that are biased towards the season of measurement. In addition discarding the incomplete data represents a significant waste of time and effort in the research study. A work around to reduce the bias problem is to predict the missing data from other measured variables using machine-learning techniques. Some questions that follow are: How much data is necessary to be able to train a regression model? What is the expected error of such prediction? What is the best model for such a task? This paper addresses the problem of completing a data set for the interior temperatures inside a passive house using different monitored predictors such as exterior temperature, humidity, wind speed, visibility, pressure and electrical energy use inside the building. Two regression models, multiple linear regression and random forest are compared using learning curves for the training and testing sets for visualizing the so-called bias-variance trade off. The learning curves help to answer the question of optimal sample size for training, model selection and expected error. Finally, descriptive statistics such as median, maximum, minimum, and room temperature averages are presented before and after completing the data sets.

Keywords: Learning curves, multiple linear regression, random forest, passive house, temperatures, sample size