The impact of weather and manual airflow management on indoor CO2 concentration

Olga Terekhova*
Mothers and Machine Learning Course Capstone Project
Vector Institute
Toronto, ON
olga.terekhova@gmail.com

Executive Summary

We demonstrate how weather factors and manual airflow management (opening windows) can be used in predicting CO2 levels in an apartment. We use data with weather measurements and indoor Smart Home solution logs and we build a machine learning model to predict CO2 levels and interpret the importance of factors available in data. We see that both sets of factors impact predictions, that's why they should be taken into consideration when trying to improve indoor ventilation.

1 Introduction

COVID pandemic highlighted the importance of good indoor air quality and ventilation. The problem is especially pressing in schools where there is a lot of people concentrated in classes. Many school buildings are old and don't have adequate mechanical ventilation. In the context of COVID pandemic the efforts were concentrated on filtering the air with HEPA filters and upgrading ventilation systems in school. During winter 2022 some teachers also tried to resort to manual airflow management by opening windows in classes, which caused temperature in classes to drop and started a debate whether opening windows is useful at all.

A study Investigation of Indoor Air Quality Booth [2021] performed in three Toronto schools measures ventilation quality by indoor CO2 levels and claims that opening windows doesn't help with ventilation:

? Did having windows open in the classrooms lead to improved ventilation?

- No. While opening windows did not help significantly reduce CO2 concentrations during the study, it is possible that small improvements in ventilation could be obtained through a more systematic pattern of window operation.

? Did the size of the window openings impact ventilation rates based on the number of open windows and CO2 concentrations in the classes?

- No. No clear benefit was evident by opening windows.

These results are puzzling. In the author's household a DIY Smart Home system is implemented which measures CO2 concentration in the apartment and notifies the inhabitants when the levels go higher than 700ppm. After the notification the inhabitants open windows wider until the levels drop down below threshold. Intuitively it seems that the system works well and demonstrates that opening windows leads to better indoor ventilation.

*LinkedIn: https://www.linkedin.com/in/olga-terekhova/
GitHub: https://github.com/olga-terekhova

Preprint. Under review.

The goal of this paper is to measure the effect of opening windows quantitatively and to explore the role of different factors in indoor CO2 levels. More specifically we are interested in finding out:

- whether outside weather impacts indoor CO2 levels (it might be harder to manage them when it's very cold or hot outside)
- whether manual airflow management (opening windows) influences indoor CO2 levels
- what is the predictive power of available factors.

2 Problem definition

We have a goal to understand better how to improve indoor ventilation (as measured with CO2 concentration by proxy). To do this in an ideal scenario we would measure the impact of all possible factors and give our recommendations based on it. Gathering data for this purpose would be time and resource consuming and should be done in subsequent studies designed to explore the problem area further. In this paper we'll concentrate only on factors available in data:

1. outdoor weather measurements in Toronto
2. logs of indoor measurements provided by the DIY Smart Home system in the author's household (located in Toronto).

We'll measure the impact of the available factors on CO2 concentration levels.

This data is stored in the following datasets:

1. The data with **outdoor** weather metrics is sourced from Environment and Climate Change Canada. We take daily readings from the last 1000 days. For the purpose of our paper we are interested in 7 attributes: date, average temperature, average humidity, average dew point, average wind speed, average pressure, average health index, precipitation. Original download link: `https://toronto.weatherstats.ca/download.html`. The resulting dataset is hosted at `https://raw.githubusercontent.com/olga-terekhova/weather-co2/main/data/weatherstats_toronto_daily.csv`.
2. The data with **indoor** measurements is sourced from the Smart Home system's log, aggregated to hourly readings. There are three sensors in different locations in the apartment and the data averages their readings. The data has 5 attributes: date, hour, average indoor CO2, average indoor humidity, average indoor temperature. The period of observation is from March 2021 to April 2022. The resulting dataset is hosted at `https://raw.githubusercontent.com/olga-terekhova/weather-co2/main/data/indoor_stats_hourly.csv`.
3. The final dataset for this study is produced by merging the two source datasets together and contains features with both **outdoor and indoor** measurements. The source datasets are joined on the 'date' attribute. The final dataset is created as part of the current study in a Colab notebook.

Our approach is to run machine learning models on the final dataset to measure the predictive power of the features in the dataset and to interpret the importance of these features for the impact on predictions.

For our task we evaluated 5 regression machine learning models that predict a continuous variable 'indoor CO2'. See Table 1:

For our problem we consider the Random Forest model to be the most suitable one. It produces good score for prediction using input features and it gives some interpretability using feature importance, which can be further enhanced by exploratory data analysis.

3 Model

We run a Random Forest Regressor model using scikit-learn Python package.

The dataset is split into a train and a test datasets in the proportion of 90% to 10% rows respectively. There are nine input variables:

Table 1: List of evaluated ML models:

Name	Score	Interpretability	Fit for the problem
Linear Regression	7.7%	Limited	Data is not linearly separable, so the model doesn't produce good results
Decision Tree	41.8%	Good	Decent predictive power and interpretability (feature importance and tree visualization), although at depth = 20 it's hard to interpret the tree visualization
Random Forest	68.4%	Limited	Good predictive power and some interpretability using feature importance
Neural Network	4.4%	None	Might not be very suitable for the kind of tabular data in the study, might need too much additional effort to finetune architecture and hyperparameters to get decent score
Nearest Neighbours	77.7%	None	Although the performance is the best among all models, the model doesn't give us interpretability.

1. hour : hour of the measurement (0-23)
2. month : month of the measurement (1-12)
3. dayofweek : day of the week (0 - Monday, ..., 6 - Sunday)
4. avg_indoor_temperature : average indoor temperature
5. avg_hourly_temperature : average outdoor temperature
6. avg_hourly_relative_humidity : average outdoor relative humidity
7. avg_hourly_wind_speed : average outdoor wind speed
8. avg_hourly_pressure_station : average outdoor pressure
9. avg_hourly_health_index : average outdoor health index

The output variable that the model is trying to predict is CO2 concentration (avg_indoor_co2). All variables involved are numerical and no additional encoding is needed. There is no multicolinearity for independent variables chosen.

Training of the model is performed on the training dataset and the model is tested on the testing dataset. We use squared error as the criterion for the model and we choose the best depth for the forest from the list of values: 2, 5, 8, 9, 10, 15, 20, 30, 50, 70.

The model is implemented in a Colab file at `https://colab.research.google.com/drive/1utrxtvtDamlRplIsliJq9Gh4VsDX3B2B?usp=sharing`.

The code for the model is partly based on materials from Vector Institute Institute [2022] and Stackoverflow Stackoverflow [2022].

4 Results and findings

The best model is the Random Forest Regressor at the depth = 30. Its score on test dataset equals 68.4%, which is quite good. We may conclude that the features selected for the model explain more than half of variation of CO2 concentration in the selected household.

To understand what factors are the most impactful for prediction we look at the feature importances for the model. See Figure 1

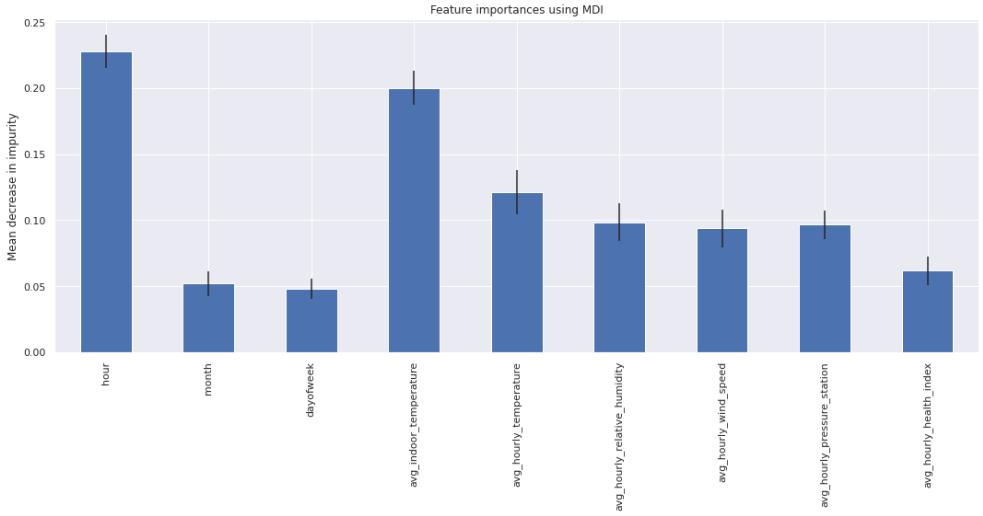

Figure 1: Feature importances

The most important features were hour and average indoor temperature. They may be seen as manual airflow management factors.

As seen on Figure 2, the CO2 concentration is higher at night, when the inhabitants don't manage airflow, lower in the morning, when they react to notifications and widen the window gaps, higher in later morning (maybe concentrated at work and not noticing the notification timely), lower in the evening and the lowest at bedtime, when the apartment is aired out to cool down and ventilate the bedroom.

As seen on Figure 3, the levels of CO2 are the lowest at the lowest indoor temperature levels. That may mean that opening the windows lowers both CO2 and indoor temperature. It might also mean that the movement of the air inside the unit matters. As the thermostat in the unit is set at around 23, higher indoor temperature and higher CO2 levels observed might mean that there is a cool pocket near the window and the thermostat, and the ventilation inside the unit doesn't equalize levels of CO2 and temperature in the whole unit properly.

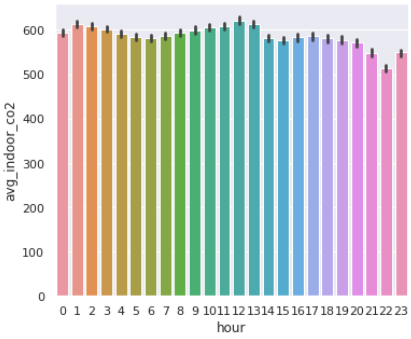

Figure 2: Relationship between hour and indoor CO2

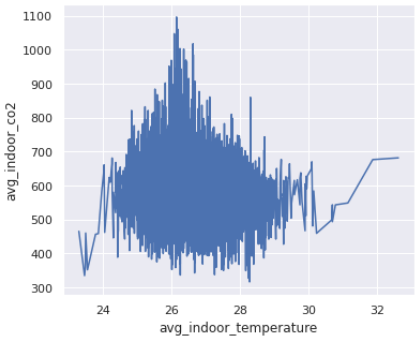

Figure 3: Relationship between indoor temperature and indoor CO2

It's also important to note that the CO2 variable appears to be normally distributed with the mean at 600ppm (as seen on Figure 4). It might mean that the current notification system and the processes of managing airflow by opening windows works quite effectively.

The outside weather factors do not have the same impact as manual indoor airflow management, but they do however explain some variation in CO2 levels. The impact is not very strong and the results do not support the hypothesis that it might be more difficult to manage ventilation when it's very cold or hot outside.

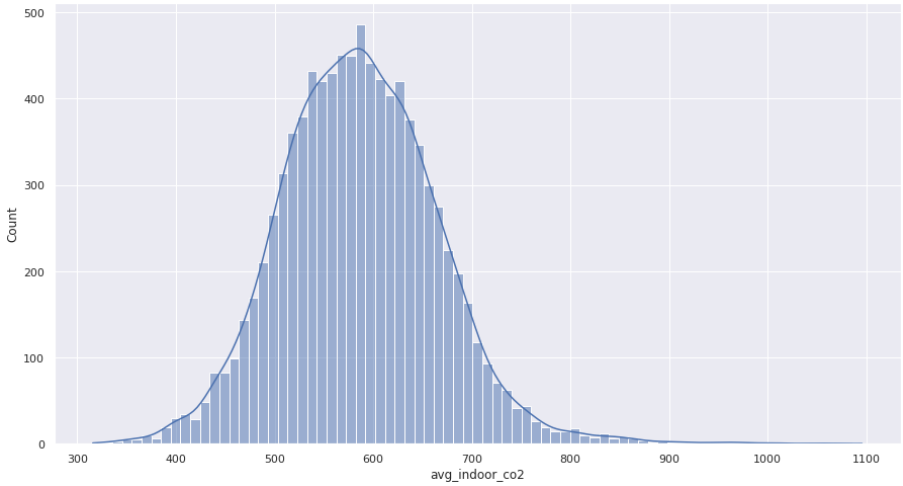

Figure 4: CO2 levels distribution

The results clearly show that weather and manual airflow management factors impact CO2 levels, even though a lot of variability is explained in factors omitted from this study (like the number of room occupants).

5 Conclusions and future work

The results of our study show that opening windows should bring down CO2 levels and it should be considered as effective technique to improve ventilation in households and organizations.

Further studies are needed to explain the difference of conclusions compared to the study in Toronto schools Booth [2021]. Limitations of the current study are:

- confounding of weather and manual management factors - it might be useful to examine them separately while controlling the other set of factors,
- only one household was examined with a very specific ventilation profile - a Smart Home system implemented, heating and cooling combined with ventilation system in the unit, an additional fan installed near the balcony door, powerful bathroom fan turned on sometimes, some air coming from the lobby through door cracks. So the results might not generalize well to other households and organizations,
- there were no measurement of other obviously relevant factors like the number of people present in the unit, whether a bathroom fan is on, or the width of window gaps.

For future studies other types of households and organizations should be assessed as well as more factors should be measured.

In the context of the approach chosen for the current study, future work might be the following:

- using the weather dataset aggregated to hourly rather than daily measurements - it might show stronger relationship between indoor hourly CO2 levels and outdoor weather factors,
- measuring additional factors in the household, like whether the bathroom fan and the additional fan near the balcony door are turned on - it might show the effect of internal air circulation in the unit.

6 References

Stephen Booth. Investigation of indoor air quality. Technical report, Pinchin, February 2021. URL https://tcdsbpublishing.escribemeetings.com/filestream.ashx?DocumentId=22656.

Vector Institute. Code assignments for Mothers and Machine Learning course, April 2022.

Stackoverflow. Stackoverflow code examples, 2022. URL https://stackoverflow.com/.