Course Project
Estimation & Optimization of building operational energy emissions
29/11/2023

Contents

1. Preface

This project is inspired by a study conducted by the Swiss Federal Statistical Office for estimating emitted CO_2 of residential buildings by using climate and energy-relevant information [1]. As described in the course project requirements, I have no information about the exact approach of the organization. My goal is not only to infer what kind of analytical models might have been used in this particular case, but also to suggest a more general solution for estimating emissions for buildings that can be partially residential, such as office buildings, hotels, schools, kindergartens, etc. I then propose a prescriptive analytics solution to optimize building parameters to minimize emissions given constraints such as cost and other resources.

This topic reflects a challenge that I have been thinking about for a while but have not had the analytical tools to explore. It's a combination of multiple business cases and research projects that I have been exposed to in the last several years. I would like to use this as an opportunity not only to demonstrate what I have learned

throughout the course, but also to introduce a topic that has yet to be revolutionized by data science and analytics.

2. Introduction

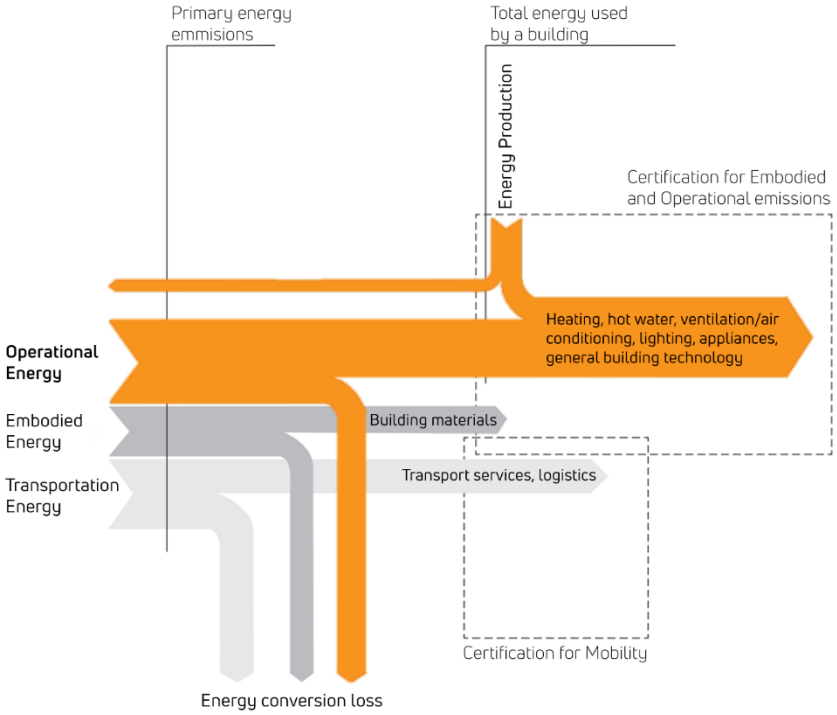


Figure 1: Energy flow from primary energy to different types of usage use [2]

Globally, building operational emissions (Figure 1), account for about one-third of all CO_2 emissions worldwide [3]. In this context many governmental organizations, primarily in North America and Western Europe, are taking regulatory measures to certify buildings based on their ecological impact. These regulations highly impact both municipalities and real estate developers. Owning a certified sustainable building is becoming more than just an aspiration for a green corporate image, it is becoming a mandatory requirement.

To calculate building CO_2 emissions, real estate developers, investors, construction managers, and many others in the architecture, engineering, and construction industry turn to sustainability experts who are well acquainted with standards and know how to implement all the rigorous calculations. However, assessing a building's CO_2 impact based on energy performance is rarely carried out due to its lengthy and costly nature. This is a problem because sustainability calculations are naturally pushed to the end of the project design phase when changes are very limited and costly. In order to save time and manpower, it has become essential for investors and project managers to be able to predict whether the desired standardization requirements will be met for a particular project at an early design stage. By considering a few building characteristics and using them as input for an analytical model, it is possible to estimate building operational emissions before the entire certification process is carried out. What's more, the prediction can then be used as an objective function to optimize building parameters to find the optimal solution.

One way to estimate building performance based on a given standard is to measure it against similar buildings that have already been certified and one country that is currently a pioneer in sustainable development is Switzerland. There, there are hundreds of thousands of certified buildings and their certificate metrics are available online. In addition, online public registers provide easy access to building characteristics such as location, gross area, number of floors, footprint geometry, etc. On top of that, other public geospatial

resources, such as local weather data from climate stations and emission coefficients for energy resource types, can provide valuable insights into how a building performs (Figure 2).

Figure 2: Open-access certification and geo-spatial data provided by the Federal Authorities of the Swiss Confederation © swisstopo, SFOE [4]

3. Methodology

In this course project it is assumed that information about certified buildings, including operational CO_2 emissions, from sustainability agencies, building characteristics information from municipalities and geo spatial data sets from geographical information platforms is provided. The main goal is to use supervised learning to predict operational energy CO_2 emissions for a new building project given basic building characteristics, such as location, function, year of construction, structure type, façade materials and geometry. Then use the predicted emissions as an objective function to optimize building parameters such as energy providers, construction type, % glazing, etc. (Figure 3).

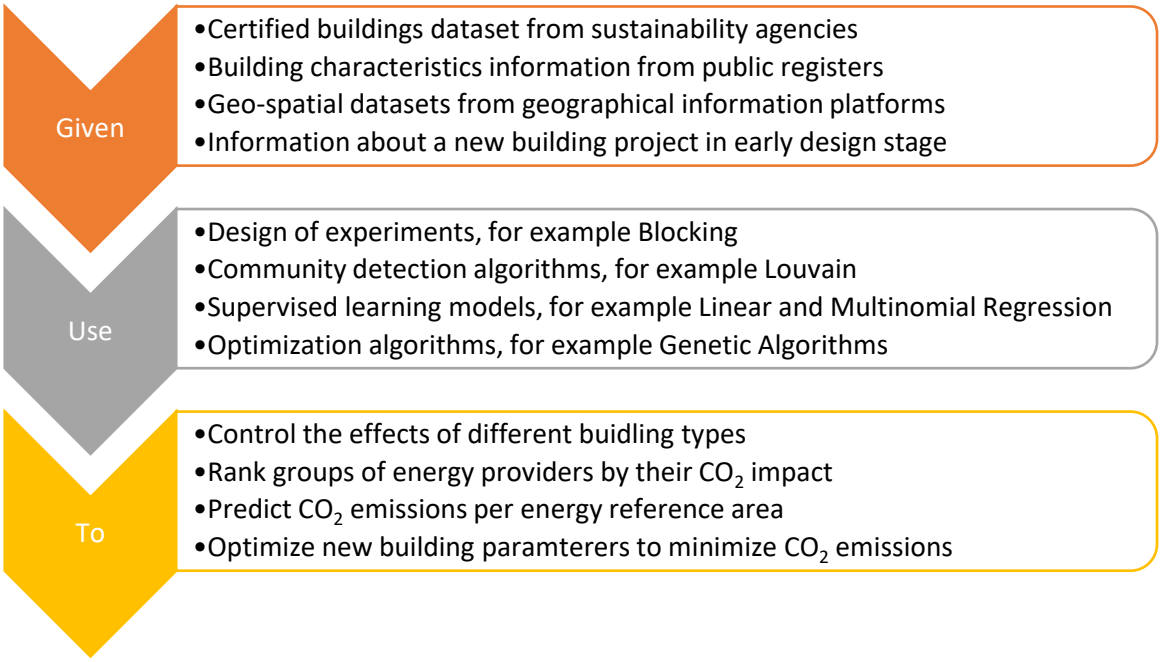


Figure 3: A summary of the {Given}{Use}{To} approach used in this project

3.1. Overview

This section is organized as follows. Part 1 introduces the problem setup. Part 2 covers what alternatives there are. Part 3 shows the different models and how they are linked. Part 4 describes the different data sources and their relationships.

3.1.2. Controlling the effect of different building categories

The calculated emissions are an expected value for a specific building category. In this case, Blocking can be used to rearrange buildings that are similar to one another in groups [5]. Thus, Explanatory Variables are the buildings characteristics, the Response Variable is the amount of CO_2 emissions and the Nuisance Variable would be the building type (Figure 4). By doing this, the variation within each block would be much lower compared to the variation among all buildings. A better understanding of how different combinations of energy sources affect Operational Energy CO_2 emissions would be gained while controlling for building category.

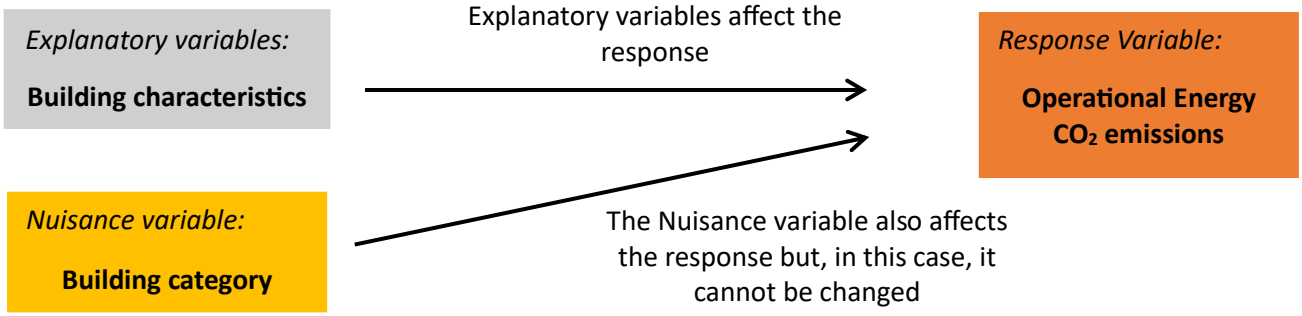


Figure 4: Blocking strategy used to group by building type

3.1.2. Community Detection for grouping energy providers

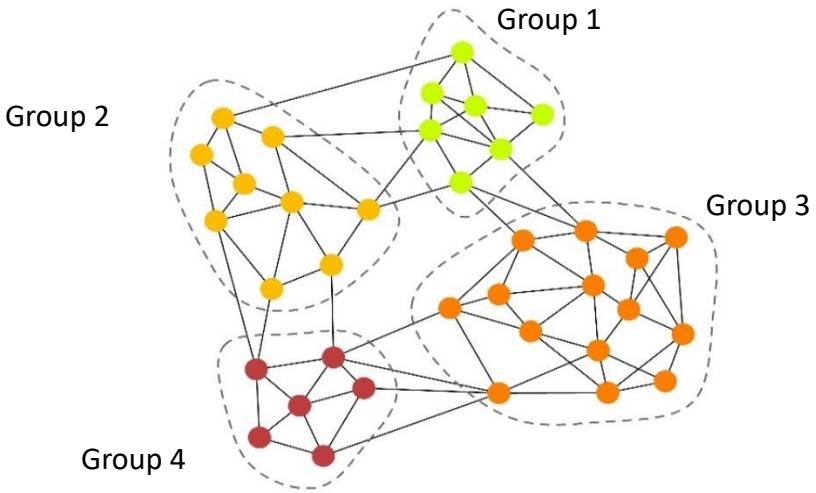


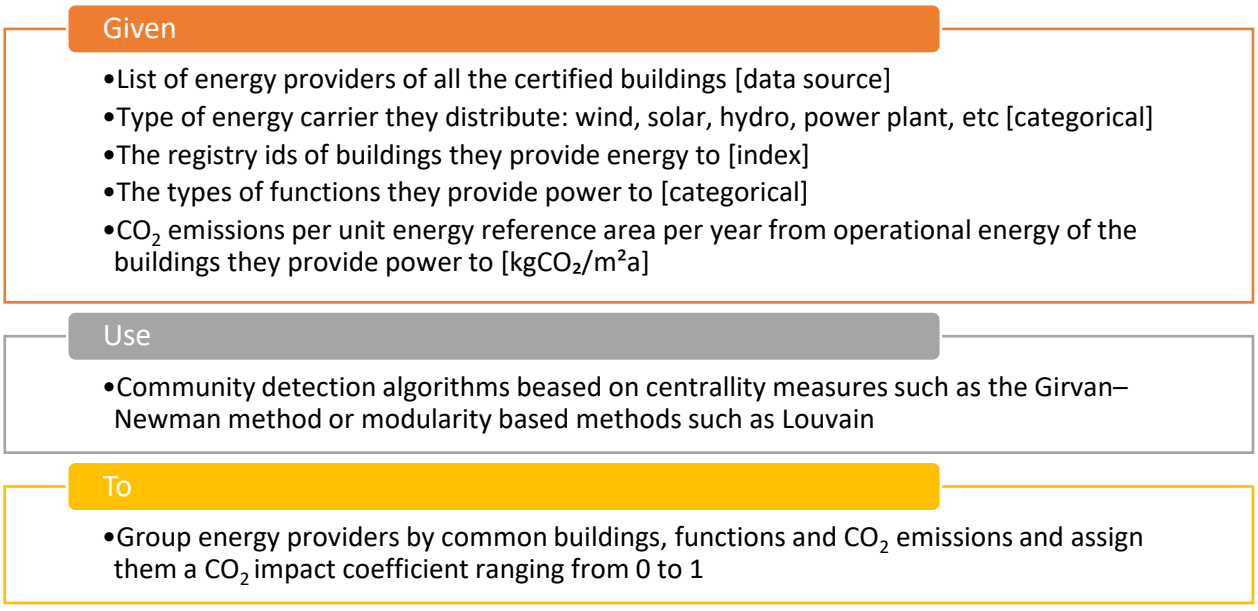
Figure 5: Grouping energy providers (nodes) by the buildings (edges) they power, then ranking groups by their CO_2 impact (green – low, red – high) using community detection algorithms [6]

In practice, total energy demand needs to be calculated before calculating CO_2 emissions. This can be very time-consuming since the total energy demand of a building is the sum of energy required for heating, cooling, ventilation, lighting, appliances, warm water and other functions. CO_2 emissions vary greatly depending on the energy source for each of these functions and each function can have one or more different energy sources from different providers (Figure 6). Furthermore, energy demand calculations are also weather adjusted, as annual temperature variations and elevation can lead to large fluctuations. Therefore, calculating the effects of all these factors and comparing the results can be inefficient in terms of time and resources.

	heating	hot water	ventilation	appliances	lighting	facilities	other
Fuels	✓	✓					
Teleheating [7]	✓	✓					
Power grid [8]	✓	✓	✓	✓	✓	✓	✓

Figure 6: Table showing which functions can be powered by which energy carriers

Furthermore, energy carriers can have different providers based on the building's location. One solution is to implement a sustainability index group ranking system by finding which energy providers are strongly related to one another in terms of powering the same building. Because data of certified buildings is already present, energy carrier providers can be related to the respective amount of operational emissions. That way, after finding the groups, they can be ranked by their CO_2 impact. Then, for an uncertified building, after specifying its energy providers for each type of function we can see in which group they are and use the average sustainability index of the groups (Figure 5).



3.1.3. Simple Regression for energy reference area estimation

The energy reference area (ERFA), also known as the heated gross floor area (GFA), is the sum of all the floor areas of a building that are heated or air-conditioned. In most of the certification requirements CO_2 emissions are measured per energy reference area. However, in the early design stages rooms are not yet fully defined

and neither their area nor their type is known. Given data from sustainability agencies for ERFA and other building characteristics data provided by geographical information platforms, Lasso Regression can be used for variable selection. In this case it can be assumed that variables don't seem to be dependent on one another. Thus, Simple Linear Regression with or without interaction terms to would be appropriate to predict ERFA:

Given

- Building type (control the effect as proposed in 3.1.1) [categorical]
- Building age [years]
- Building gross area [m^2]
- Building footprint area [m^2]
- Building elevation (from geo location) [m]
- Number of occupants [integer]
- Number of rooms [integer]
- Attic area [m^2]
- Utilities area [m^2]
- Mean early outdoor temperature at building location [C°]

Use

- Lasso Regression or Elastic Net for variable selection and Simple Linear Regression with or without interaction terms (e.g. Number of floors = gross area / footprint area)

To

- Predict building energy reference area [m^2]

3.1.4. Advanced Regression for estimating CO_2 operational emissions

Unlike point 3.1.3 in this case there are interactions between the variables which we do not know. This is why advanced regression models would be more appropriate. The following set-up is proposed:

Given

- Building type (control the effect as proposed in 3.1.1) [categorical]
- Building elevation (from geo location) [m]
- Mean early outdoor temperature at building location [C°]
- Building energy reference area (estimated in 3.1.3) [m^2]
- Building floors count [integer]
- Building footprint area to perimeter ratio [number]
- Building envelope glazing percentage [number]
- Building age [years]
- Building energy providers sustainability group index (estimated in 3.1.2) [number]
- Building renewable energy production per year (e.g. photovoltaics) [kWh/a]
- Building structure (wood, concrete, steel) heat capacity (low, medium, high) [rank]

Use

- Lasso Regression or Elastic Net for variable selection and Gradient Boosted Random Forest, Regression Splines or K-Nearest Neighbors Regression

To

- Predict CO_2 emissions per unit energy reference area [$kgCO_2/m^2a$]

Heating and ventilation are some of the biggest consumers of energy in buildings. Heating demand is a function of transmission heat loss, ventilation heat loss, solar heat gain, and internal heat gain. Transmission heat loss is mainly influenced by the thermal transmittance of the building envelope and its total area. Thermal transmittance is highly correlated with how long ago a building was renovated or constructed. Total building envelope area is highly correlated with building gross area and number of floors. Ventilation heat loss is highly correlated with the building type – a restaurant would require a greater amount of air exchange per hour per unit area than a single-family home. Solar heat gain is highly correlated with percentage of glazing. Internal heat gain is also correlated with the building type, e.g. occupant count per unit area and electricity demand.

3.1.5. Optimizing building parameters to minimize CO_2 emissions

In this step it is assumed that the building footprint geometry is parametrized. Thus, footprint area to perimeter ratio is determined by the desired footprint geometry (Figure 7).

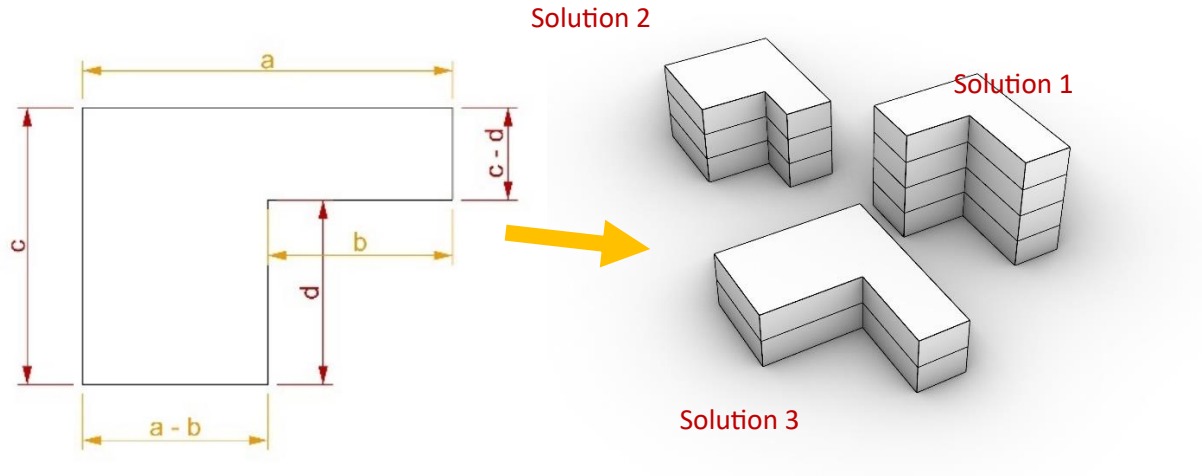


Figure 7: Building footprint parametric model for "extruding" 3D geometry

Here is an overview of the optimization model set-up:

Variables

- Renewable energy production per year calculated from m^2 of photovoltaics [kWh/a]
- Energy providers for each function: heating, ventilation, etc. [index]
- Building envelope glazing percentage [number]
- Building structure (wood, concrete, steel) heat capacity (low, medium, high) [rank]
- Number of floors [integer]
- Parameters controling the building footprint geometry (a,b,c,d in Figure 7) [number]

Constraints

- Project budget [number]
- PV should not exceed available facade/roof area [m^2]
- Minimal and maximal glazing % of facade [number]
- Minimal and maximal project building footprint area [m^2]
- Minimal and maximal project building gross-area [m^2]
- minimal and maximal number of floors [integer]

Objective Function

- Minimize CO_2 emissions using model from point 3.1.4 [$kgCO_2/m^2a$]

It is evident that this optimization model falls into the category of general optimization models. At this stage it must be considered whether a heuristic approach such as this one would give satisfactory results.

3.2. Model Selection

In analytics modelling trade-offs such as accuracy-simplicity, speed-cost, bias-variance and explainability - performance (also known as “know why or be right”) should be considered before final models are selected. Simplicity is key – often a very complicated model would perform only slightly better compared to a simpler one. An overly simple model will likely have high bias and low variance (underfitting) and an overly complex model would tend to have low bias and high variance (overfitting). Also, it is worth mentioning that sometimes doing nothing is the best solution in terms of added value, resources, profit and other metrics.

3.2.1. Alternatives to using Blocking for building category

A much simpler approach may be to just split our data into different sets, each for every different building category. However, this would be sufficient only if enough data points are provided. There is trade-off between sampling a bigger data set and solution simplicity. A thorough analysis must be made to see whether this step would improve overall model performance.

3.2.2. Alternatives to Community Detection for energy providers

One can argue that clustering is also a good approach in this case. Like community detection, clustering is a machine learning technique that groups similar data points into the same cluster based on their attributes. Although clustering can be applied to networks, it is a broader field of unsupervised machine learning that deals with multiple attribute types. On the other hand, community detection is specifically tailored for network analysis, which is closer to the case of common energy providers for buildings. In addition, clustering algorithms tend to separate individual peripheral nodes from the communities to which they should belong.

Finally, it must also be evaluated if this particular step is actually needed. The service providers may also be examined only as categorical variables for each building function (heating, water, ventilation, etc). An even simpler approach would be to take only the provider that contributes to the largest proportion of energy and it also might be possible that most buildings have only one energy provider.

3.2.3. Alternatives to Simple Regression for energy reference area prediction

Data about energy reference area (ERFA) to gross area ratio can be calculated from the buildings’ certification data. It might be more efficient to use the mean ERFA to gross area ratio for each building category. For example, it might be that the ERFA of an office building is 70% while the ERFA of a residential building is 85% of the gross area. Here there is a trade-off between accuracy and performance and it must be decided which one is more important in terms of added value.

3.2.4. Alternatives to Advanced Regression for CO_2 emissions per ERFA prediction

Although there are many advantages to using advanced regression models such as Gradient Boosted Random Forest and Regression Splines, very big attention must be given to avoiding overfitting. The more “curved” the hypersurface of the regression models is the more it can “twist” itself to fit random patterns in the data.

One obvious alternative would be to use simple linear regression. It may not be as accurate as the models proposed in 3.1.4 but it would make the overall solution much simpler and easier to interpret. Here there are trade-offs between performance-accuracy and accuracy-interpretability.

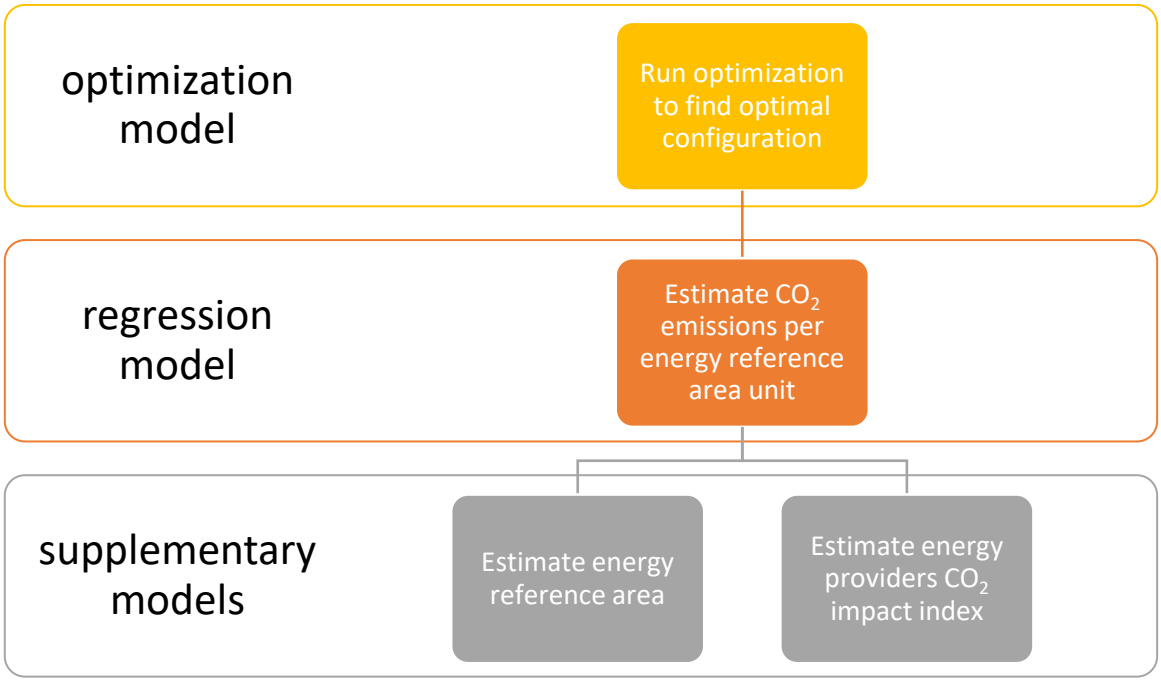
3.2.5. Alternatives to using Optimization algorithms for building parameter optimization

General optimization models such as Genetic Algorithms can be very versatile in terms of performance and results. However, optimization models assume that all the values of the input data are known exactly, which in most construction projects is not true. For example, project budget might suddenly be much less than expected – then the optimization process must be run again but with updated constraints.

The alternative at this point is to not use optimization at all – estimates of CO_2 emissions for multiple manually configured project set-up might be just as useful for the design teams. In this case it is a question of use case scenario and user studies.

3.3. Model interactions

Assuming that all models remain the same after considering all points in 3.2., this section provides a visual representation of how the different models interact with one another:



3.4. Data sources, integrity and validity

This project proposes the use of three different data sources: building certification data, building public registry data and administrative geo-spatial data sets (Figure 8) To gain as much information as possible, a strategy for "intersecting" these data sets by common features such as building registry id must be implemented.

building certification data	building public registry data	administrative geo-spatial data sets
•building regestry id •buillding category •building energy reference area •building facade glazing % •building energy demadper year [kWh/a] •building heating demadper year [kWh/a] •renewable energy production per year [kWh/a] •total CO2 per year [$kgCO_2/a$] •CO2 per unit energy reference area per year [$kgCO_2/m^2a$] •operational total CO_2 per year [$kgCO_2/a$] •operational CO_2 per unit energy reference area per year [$kgCO_2/m^2a$] •embodied total CO_2 per year [$kgCO_2/a$] •embodied CO_2 per unit energy reference area per year [$kgCO_2/m^2a$] •**many more...**	•building regestry id •building category •building sub category •building address •building city •building owner •date of construction •building gross area •building site area •buildign floors count •building energy sources •building energy providers •**many more...**	•Buildings: •regestry id •category •address •geo location •footprint geometry 2D •volume geometry 3D •Weather stations: •address •city •geo location •mean yearly temperature •mean monhtly temperature •elevation •Energy providers •energy carier type •buildings provided with energy •building functions powered •cities powered •**many more...**

Figure 8: Different data sources and what types of information they provide

An important question would be how much data is needed. Assuming that all data sources mentioned so far are open access, the answer is – as much as possible. One realistic number would be 100 000 buildings total. However, more data might be needed from the sustainability agencies and they might require some form of payment and this must also be taken into consideration.

It should also be noted that data records may not always match in terms of structure. For example, a building certification record from 2000 would likely be very different from a one from 2020. A "harmonization" methodology must be applied to ensure all data comes in the same formatting.

One must also ask: Will there be outliers? For example, buildings that are very uncommon (Figure 9) may have much worse performance in terms of CO_2 emissions but may still have been certified under some very rare conditions. A decision must be made if those buildings should be included in the training data or removed.

Figure 9: Example of a building outlier (Image source: [9]*)*

Also, a strategy for dealing with missing data must be devised. A “rule of thumb” is to use imputation if missing data is less than 5% per variable. Otherwise, other variables for indicating missing data must be introduced or new data must be gathered to fill the gaps.

Finally, it should also be planned how often to update the data and retrain the models. Standards change frequently and most of the certified buildings adhere to older versions of the standards. Perhaps correction terms have to be introduced in the short term and every six to twelve months the models should be retrained and reevaluated.

4. Models performance and maintenance

Establishing performance criteria before modeling begins is critical [10] . If this course project were to reach production its performance should be measured constantly and some form of change detection in performance metrics should be implemented. Another important task is to compare model predictions based on newly certified buildings.

One thing that would cause the models to degrade is a change in the current situation. For example, it might be possible that in a few decades most or all energy providers would start distributing only renewable energy. Then, the entire modeling process must be redone – the CO_2 prediction model would need to be retrained or it might even be deprecated since there may no longer be any CO_2 emissions to predict.

Another reason would be a change in the data stream. For instance, certificate requirements and calculations may be entirely different in a few months. The solution would then be to refit parameters and analyze the models’ performance.

5. Conclusion

There are certainly many other approaches to solving the problem of estimating building operational CO_2 emissions. In this project I have shared my own thoughts and combined what I’ve learned throughout the course. Will my approach be applicable in real case scenarios? Based on my observation, probably yes, but I cannot completely escape from my personal biases and assumptions. In order for this approach to be reliable it must be thoroughly validated both from analytics professionals and sustainability experts.

References

[1] "Interactive CO2 calculator on the Swiss federal government's geoportal," [Online]. Available: https://www.bafu.admin.ch/calculator-co2-buildings.

[2] "Overview of swiss standards for sustainable buildings," [Online]. Available: https://map.arch.ethz.ch/artikel/38/hilfsmittelgebaudelabels.

[3] "Bringing Embodied Carbon Upfront report," [Online]. Available: https://worldgbc.org/advancing-net-zero/embodied-carbon/.

[4] "Maps of Switzerland federal geoportal," [Online]. Available: https://s.geo.admin.ch/xu1lk8eay44v.

[5] "Blocking in Statistics: Definition & Example," [Online]. Available: https://www.statology.org/blocking-statistics/.

[6] "Communities detection with NetworkX," [Online]. Available: https://networkx.org/documentation/stable/reference/algorithms/community.html.

[7] [Online]. Available: https://en.wikipedia.org/wiki/District_heating.

[8] [Online]. Available: https://en.wikipedia.org/wiki/Electrical_grid.

[9] [Online]. Available: https://www.architectureanddesign.com.au/features/list/weird-architecture-world-s-top-10-weirdest-buildin.

[10] C. Kozyrkov, "Setting performance criteria for AI," [Online]. Available: https://youtu.be/lCwzCBM5WoY?si=FPXzt10tfyBG24Dm.

**Software used for Figures 5 and 7 2D/3D sketching: Rhino 3D*