Estimating a Swiss Building Stock Model

Conceptual Design Report

Mathematical Institute

University of Bern

Supervisor

Dr. A. Mühlemann

Student:	Gilbert Franz Böhme
CoS:	CAS Applied Data Science
E-Mail:	gilbert.boehme@students.unibe.ch
Date:	05.10.2025

Abstract

This report presents the conceptual design of an updated building model developed to represent and forecast the evolution of the Swiss building stock. The model was originally established to support the Federal Office of Energy in deriving energy reference areas and has since been revised to incorporate more recent data sources such as the population and housing censuses, insurance databases, and construction investment statistics. The primary objective of the redesign is to provide a consistent, transparent, and practical framework for estimating building volumes and energy-related floor areas across different periods, building types, and uses. The resulting framework provides a robust basis for monitoring the structural and energetic characteristics of the building stock. In addition, it provides essential input for energy policy, urban planning, and sustainability strategies. This conceptual framework establishes a valuable tool for decision-makers in both the public and private sectors.

Contents

1 Project Objectives 4

2 Methods 5

3 Data 6

4 Metadata 7

5 Data Quality 8

6 Data Flow 8

7 Data Model 10

8 Documentation 11

9 Risks 11

10 Conclusions 13

References 14

A Appendix 15

1 Project Objectives

The objective of the project is to estimate the volume and floor area of the Swiss building stock in a time-series context, as no public dataset currently provides this information. Currently, the Federal Statistical Office (FSO) provides data on the number of buildings and dwellings, but not on their volume or energy reference area. Reliable figures will be generated for both residential and commercial buildings, forming a basis for energy analyses, market studies, and policy decisions.

A secondary objective is to undertake a comparative analysis of different modelling approaches. All models will be calibrated with the most recent data to ensure comparability. The results will be systematically compared and assessed by expert. The use of expert assessments is justified by the limitations of purely statistical approaches, which are constrained by data quality and modelling assumptions. Expert knowledge adds practical experience and contextual insight, thereby increasing the credibility and acceptance of the results.

The expected outputs of the project are:

1. **Time series of energy reference area by building type and municipality** - to capture structural changes in both residential and commercial sectors over time.
2. **Model performance comparison** - to quantitatively evaluate and benchmark the predictive results of different models.

Together, these outputs will provide a robust and consistent evidence base for the analysis of the Swiss residential and commercial building stock.

2 Methods

The methodological pipeline will follow three stages:

1. Data ingestion
2. Data cleaning and structuring
3. Model fitting and evaluation

Data Ingestion

The heterogeneous data sources will be ingested in a scalable, cloud-based environment capable of supporting distributed data processing and collaborative research practices. Thus, offering a methodological advantage in that it ensures reproducibility and efficiency in the handling of longitudinal datasets of substantial size, including building and dwelling databases and construction investment statistics.

Data cleaning and structuring

During preprocessing, datasets will be cleaned and validated by removing duplicates, correcting inconsistent geographical and temporal codes, and imputing missing values with auxiliary information (e.g., interpolated construction years or insurance cross-checks). Outliers such as implausible volumes or investments will be flagged and corrected or excluded according to documented rules. Finally, the data will be harmonised by standardising variable names, units, and coding schemes across sources.

Model fitting and evaluation

In the modelling stage, several approaches will be compared: (i) a simple baseline model, valued for interpretability and methodological transparency, and (ii) advanced machine learning models capable of capturing nonlinear relationships and more complex interactions.

The simple baseline model (i) starts with last year's stock, adds new construction (excluding replacements) and any volume-increasing refurbishments, adjusts for building conversions and demolitions, and then converts from cubic meters to square meters to the energy reference area using empirical factors. The more advanced machine learning models (ii) will be chosen based on the Databricks Automated Machine Learning capabilities.[1]

To ensure reproducibility and scientific rigor, the full modelling lifecycle will be tracked with mlflow experiment logging, systematic versioning, and artifact storage.[2]

3 Data

The building stock model combines official statistics from the FSO, administrative registers, and internal documentation. The FSO Building and Dwelling Register provides structural and occupancy data for calibration and validation, while cantonal building insurance records supply volumetric information essential for stock extrapolation. Annual construction investment statistics link monetary flows to physical changes, adjusted for inflation using the FSO construction price index (reference year 1998). Finally, internal Wüest Partner AG projects—including valuations and construction plans—offer empirical benchmarks for translating between floor area and building volume. Together, these sources establish a consistent foundation for estimating the Swiss building stock in volumetric and areal terms.

[1] Please find a AutoML documentation under: https://www.databricks.com/product/automl

[2] Please find a mlflow documentation under: https://mlflow.org/docs/3.4.0/ml/

Table 1: Data sources used in the building stock model

Data Source	Description and Use
Federal Register of Buildings and Dwellings (FSO, public)	Detailed information on residential buildings and dwellings (e.g., occupancy types, housing categories). Used for calibration and validation of building stock estimates.
Building insurance institutions (non public)	Cantonal insurance data on building volumes. Basis for the extrapolation and calibration of the commercial building stock.
Construction investment statistics (FSO, accessible upon request)	Annual construction and refurbishment investments. Used to update building stock volumes by converting monetary investments into cubic meters.
Construction price index (FSO, public)	National construction price index. Used to adjust construction cost data (e.g., from the Schück database) for inflation and cost developments over time. Reference year 1998 chosen as baseline.
Wüest Partner AG valuations, measurement projects and construction plans (non public)	Internal measurement projects and building plans. Used to derive conversion factors between floor area and building volume.

4 Metadata

The building stock and household data (yearly, since 2009), combined with the construction price index (yearly, since 1998), provide the official statistical foundation. From 2000 onward, these sources are complemented by internal Wüest Partner AG project data, such as valuations and planning studies. Building volume information is currently available for 2000, 2005, and 2010, with an update request for more recent data pending. Construction investment statistics (yearly) are available from 1993 until today.

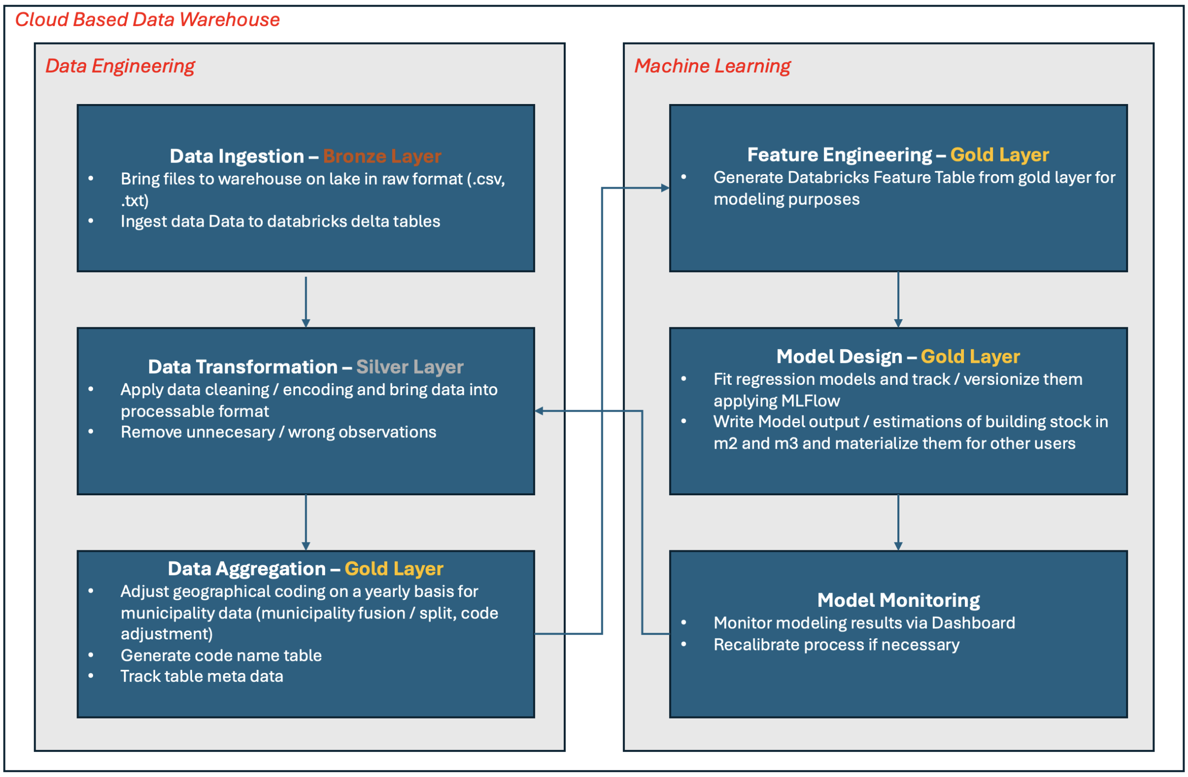
5 Data Quality

The quality of the data basis for the model can be classified as reliable. The building, household, and residential registers of the FSO are available at the municipal level as a nearly complete data basis for building data with unique identifiers (EGID, EWID, no duplicates, missing values). The data provided by the FSO on construction investments is incomplete for data protection reasons and therefore only reflects an estimate of construction investments over time. It is important to question the extent to which internal valuation data from Wüest Partner AG can contribute to modeling. Actual energy reference areas are specified in some of the valuations. Currently, 29% of all valuations contain floor and energy reference areas with an assigned building type and location.

6 Data Flow

The data processing pipeline for the Swiss Building Stock Model is implemented in a cloud-based environment that integrates data engineering with machine learning. In the data engineering stage, raw inputs such as stock data, insurance records, and construction investment statistics are ingested in their original formats (bronze layer), subsequently cleaned and standardized (silver layer), and finally aggregated with adjustments for geographical coding and metadata tracking (gold layer). In the machine learning stage, the aggregated data form the basis for feature engineering (gold layer) and model development. Regression and predictive models are trained with systematic versioning and experiment tracking, and the resulting outputs—such as building volumes and energy-related areas—are materialized for further use. The modeling results will be monitored via a tracking dashboard and quality checks in a notebook.

Figure 1: Data Flow of the Swiss Building Stock Model Input Data


Cloud Based Data Warehouse
Data Engineering
Machine Learning
Data Ingestion – Bronze Layer
• Bring files to warehouse on lake in raw format (.csv, .txt)
• Ingest data Data to databricks delta tables
Data Transformation – Silver Layer
• Apply data cleaning / encoding and bring data into processable format
• Remove unnecesary / wrong observations
Data Aggregation – Gold Layer
• Adjust geographical coding on a yearly basis for municipality data (municipality fusion / split, code adjustment)
• Generate code name table
• Track table meta data
Feature Engineering – Gold Layer
• Generate Databricks Feature Table from gold layer for modeling purposes
Model Design – Gold Layer
• Fit regression models and track / versionize them applying MLFlow
• Write Model output / estimations of building stock in m2 and m3 and materialize them for other users
Model Monitoring
• Monitor modeling results via Dashboard
• Recalibrate process if necessary

7 Data Model

At the conceptual level, the data model is designed to represent the Swiss building stock as a dynamic system that evolves over time. Relationships exist between buildings and dwellings (one-to-many), between buildings and uses (many-to-many, mediated by usage periods), and between buildings and construction activities (new build, refurbishment, demolition, conversion). The conceptual model thus captures both the static structure of the building stock and the processes that drive its transformation.

The logical data model relies on three core datasets that capture complementary aspects of the Swiss building stock and its development: the Federal Register of Buildings and Dwellings (FSO, 2025c), the Construction Investment Statistics (FSO, 2025a), and the Construction Price Index (FSO, 2025b). The building and dwelling variables provide structural and locational information at the building level (Table 2). These are complemented by construction investment variables, which record annual expenditures, backlog categories, and construction types (Table 3). Finally, the construction price index variables ensure that temporal changes in construction costs are consistently represented in the model (Table 4). Together, these datasets form the empirical foundation of the modelling framework.

A cloud-based data warehouse for data storage as well as Databricks computing capabilities represent the physical infrastructure for the project. The programming language PySpark will be used for data processing and model development, leveraging its distributed computing capabilities for scalability.

8 Documentation

The project will be documented through a structured README file, providing an overview of the data pipeline, model design, and usage guidelines. All metadata will be systematically recorded in a dedicated tracking table. The Databricks machine learning environment will handle model versioning, allowing for consistent comparisons across iterations. Finally, mlflow will be used to log and monitor summary statistics of the models, supporting traceability and performance evaluation.

9 Risks

The development of a Swiss Building Stock Model involves several risks that may affect the quality of the results, the timeline, and project costs. The main risks and mitigation strategies are outlined below:

1. Data Privacy and Inference Risks

Risk: Energy reference areas may allow indirect inferences about energy consumption and associated costs. When combined with other datasets, such as household or income data from the FSO, there is a risk that sensitive information about tenants or households could be derived.

Countermeasures: Ensure strict anonymisation and aggregation at appropriate spatial levels (e.g., municipality or district). Avoid linking with personally identifiable household or income data. Establish clear guidelines on data use and communicate transparently with stakeholders about privacy safeguards and the intended scope of analysis.

Impact: Failure to address these issues could result in legal challenges, restricted access to data, and loss of trust among institutional partners and the wider public.

2. Data Availability and Access

Risk: Key datasets (e.g., updated building volume data, insurance records, or investment statistics) may not be available in time, may have restricted access, or may contain missing values.

Countermeasures: Establish agreements with data providers and use imputation methods or alternative proxy datasets.

Impact: Delays in data delivery could extend the project schedule. Missing or incomplete data may reduce accuracy and robustness of outputs.

3. Data Quality and Consistency

Risk: Official statistics and internal project data may contain inconsistencies, coding errors, or varying definitions across years and cantons.

Countermeasures: Implement systematic data cleaning, validation rules, and cross-checks across sources. Document all adjustments in README and metadata tables.

Impact: Unaddressed inconsistencies could compromise the credibility of results. Additional cleaning may increase workload and cost.

4. Methodological Limitations

Risk: The chosen modelling approaches may fail to capture complex dynamics of building stock development, leading to biased or implausible predictions.

Countermeasures: Compare multiple model classes (regression and machine learning) and benchmark their results against expert assessments. Also adopt iterative model calibration.

Impact: Lower quality or less reliable predictions with potential longer revision cycles could extend project time.

10 Conclusions

This project demonstrates that while extensive statistical and administrative data on Swiss buildings exist, no dataset currently provides municipal-level predictions of energy reference areas. Filling this gap would deliver substantial value by enabling detailed monitoring of building stock developments and supporting energy policy, market analysis, and sustainability planning. The conceptual framework outlined in this report shows that such a dataset could be developed by combining federal registers, cantonal insurance data, construction investment statistics, and internal benchmarks.

A realistic time horizon for producing a first version of the dataset is 8 to 11 months, including:

- 1 months for securing data access agreements and setting up infrastructure,
- 4–6 months for data cleaning, harmonisation, and preprocessing,
- 2–3 months for model development and calibration,
- 1 months for validation, expert review, and final adjustments.

References

FSO. (2025a). *Construction investment statistics* [Accessed on 1 October 2025]. Federal Statistical Office. Retrieved October 1, 2025, from https://www.bfs.admin.ch/bfs/en/home.assetdetail.36182082.html

FSO. (2025b). *Construction price index* [Accessed on 1 October 2025]. Federal Statistical Office. Retrieved October 1, 2025, from https://www.bfs.admin.ch/bfs/de/home/statistiken/preise/baupreise/baupreisindex.html

FSO. (2025c). *Federal register of buildings and dwellings* [Accessed on 25 September 2025]. Federal Statistical Office. Retrieved September 25, 2025, from https://www.housing-stat.ch

A Appendix

Table 2: Relevant building dwelling variables

Variable	Type and Description
geographical_coding	int. Year of geographical coding.
region_type_code	int. Type of region code.
region_code	int. Code for region.
period_type_code	int. Type of time period code.
period_code	int. Code for time period.
egid	string. Pseudonymized building identifier, used for linking to the dwelling table.
building_category_code	int. Building category.
longitude	int. Longitude coordinate (LV95, meters). Missing coordinates imputed for full coverage. Provided only in certain cases and upon justified request.
latitude	int. Latitude coordinate (LV95, meters). Missing coordinates imputed for full coverage. Provided only in certain cases and upon justified request.
urban_quarter_code	string. Quarter code.
building_period_code	int. Building period.
construction_year	int. Year of construction.
floors	int. Number of floors.
apartments	int. Number of apartments in the building.

Table 3: Relevant construction investment variables

Variable	**Type and Description**
geographical_coding	int. Year of geographical coding.
region_type_code	int. Type of region code.
region_code	int. Code for region.
period_type_code	int. Type of time period code.
period_code	int. Code for time period.
construction_type_code	int. Construction type of building, newly constructed or modified.
construction_sub	int. Category of constructed building (e.g., detached house, apartment building).
construction_level_code	int. Level of construction, above or below ground (structural or civil engineering).
investment	bigint. Annual construction investment.
notallocated_investment	bigint. Annual construction investment not allocated on municipality level (counted at state level).
backlog_active_nextyear	bigint. Construction backlog for next calendar year (fixed).
backlog_active_total	bigint. Construction backlog total (fixed).
backlog_not_started	bigint. Construction backlog of projects approved but not yet started.
backlog_pre_approval	bigint. Construction backlog of projects not yet approved.

Table 4: Relevant construction price index variables

Variable	**Type and Description**
geographical_coding	int. Year of geographical coding.
region_type_code	int. Type of region code.
region_code	int. Code for region.
period_type_code	int. Type of time period code.
period_code	int. Code for time period.
category	string. Types of construction, covering both building construction and civil engineering.
weight	double. Weightings used in the construction price statistics, updated every five years as part of a revision.
index_value	double. Index value, calculated on the basis of October 1998 = 100.

Selbständigkeitserklärung

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Basel, 05.10.2025

(Unterschrift)