UNIVERSITÄT
BERN

Data Science Project

Cost estimation for precast columns
Conceptual Design Report

October 31, 2022

Raphael Ziegler
raphael.ziegler@gmail.com

Abstract

A prediction model for reinforced concrete columns is developed to facilitate cost estimation of construction projects. Different models are compared to find the most reliable prediction model. The analysis can be performed using software freely available on the Internet (FOSS).
The data are concrete column purchase prices from completed construction projects, which are manually entered into the source Excel file. The original values are checked for input errors as much as possible, sorted out, and serve as input for the Linear Regression Model. For the Neural Network, the data is additionally normalized so that the values are between 0 and 1. The models should reach a prediction accuracy of 70% to be used in general without further evaluation.
The analysis is performed for different data combinations to achieve the best result. Short-term market fluctuations (covid, war) cannot be considered and remain a risk in the prediction.

Contents

1 Project Objectives

This project aims to increase the cost certainty of shell construction offers and thus to be able to prepare an attractive offer for the client. In concrete terms, the aim is to predict delivery prices for precast concrete columns.
In the preliminary phase, it is challenging to obtain quotations for the estimating from the column suppliers, as they are swamped and prefer to concentrate their time on quotations that are about to be executed, which can lead to an order being received promptly. In order to obtain a price (request for quotation/forecast model), the static values, as well as the geometrical data (length, width, height), are necessary.

The column price is composed as follows:

- Column formwork (can be reused several times)
- reinforcement
- Concrete
- Labor hours for column formwork, reinforcement and concrete
- Profit

Different prediction models are compared (linear regression, neural network) to determine which model provides the most reliable results.

A reliable predictive model would speed up the estimating process and make it more reliable, as there is no need to search for similar columns in the table.

2 Methods

For this data science project, a Dell XPS 13 (see 2.1) is used with Manjaro Linux operating system.

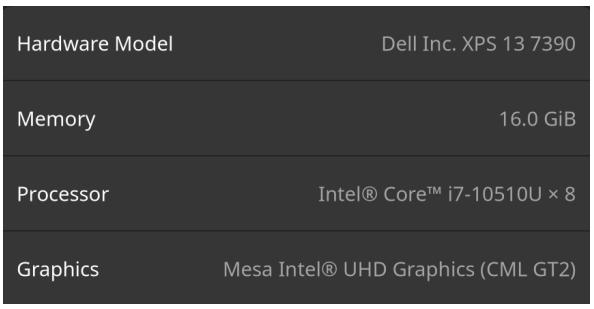

FIGURE 2.1: Used hardware

The analysis software is Jupyter Notebook [1] with Python [2]. It is planned to move the model-specific code into an extra file for a better overview in the Notebook. It is not necessary to use any online resources.

The following python modules will be used:

TABLE 2.1: Used python modules

Modul name	**Version number**
pandas [3]	1.4.4
NumPy [4]	1.23.3
matplotlib [5]	3.5.2
scikit-learn [6]	1.1.2
TensorFlow [7]	2.11.0-dev20220927

It is intended to use the linear regression model from scikit-learn and a neural network with TensorFlow for the analysis.

3 Data

The data is purchased concrete columns from completed building projects. For each purchased column, there is information about the number of pieces, geometry and column load,

further the unit prices, discount and transport costs. The table also contains calculated values, such as cross-sectional area, volume and more (see 4.1).

For data protection reasons, information such as location, project name and column supplier has been deleted or replaced by placeholders.

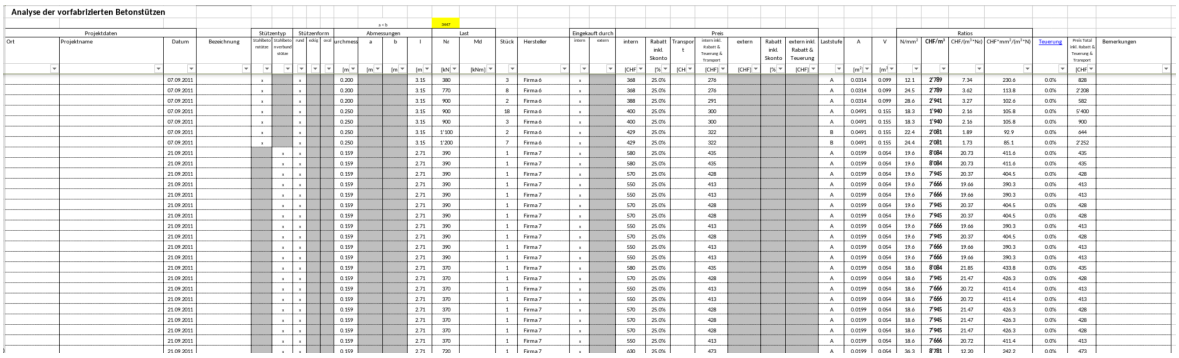

FIGURE 3.1: Extract from the column list file

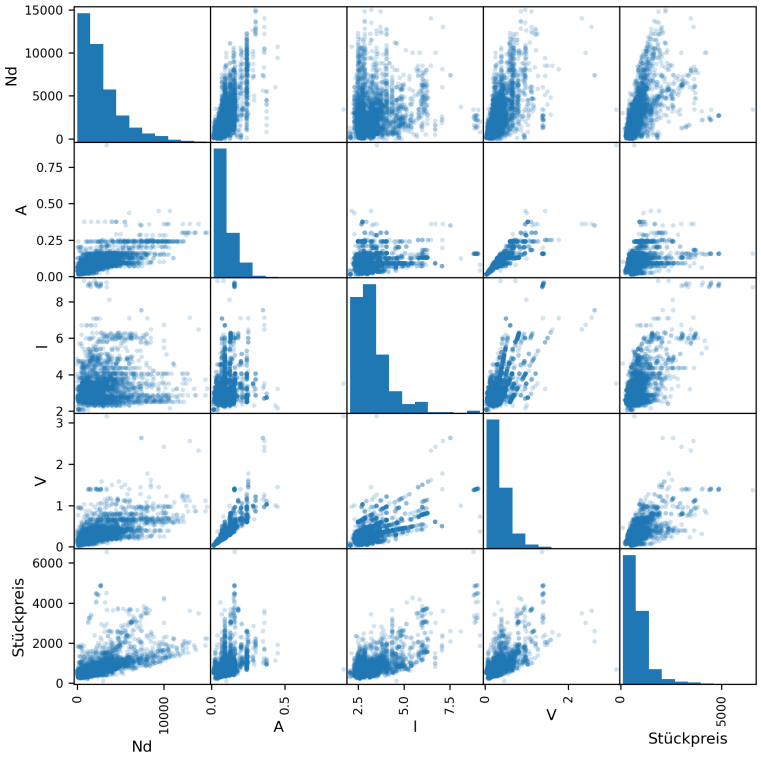

FIGURE 3.2: Scatter matrix cleaned dataframe

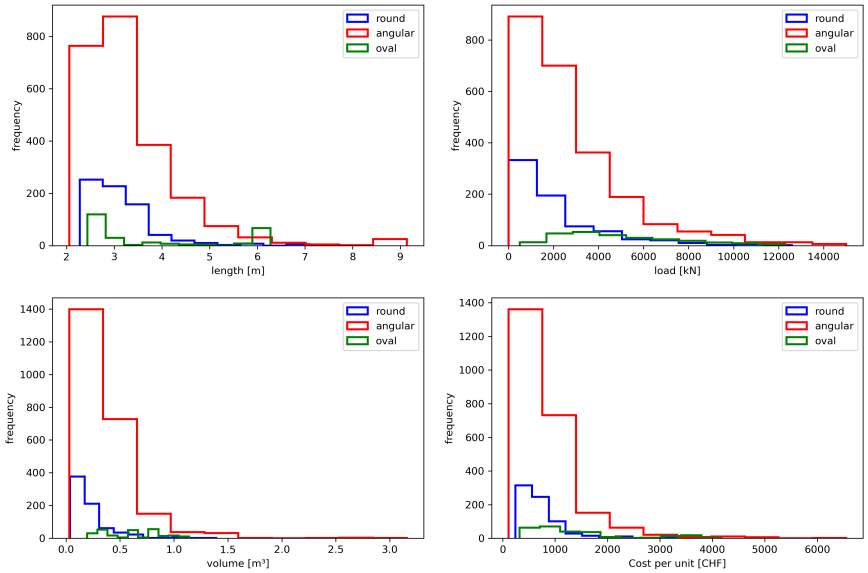

FIGURE 3.3: Histograms by cross-section types

4 Metadata

All metadata is contained in the Excel file as a value in a column or in the column's header. The table contains two different column types (see A) and three cross-section shapes.

Column types:

1. Steel/concrete composite columns: Concrete columns with steel shell and/or solid steel core. (see A.1)

2. reinforced concrete columns: Concrete columns with reinforcement (see A.2)

Cross-section shapes: Round, angular and oval (Mainly used in parking garages)

TABLE 4.1: Column names and units

Column name	**Description**	**Unit**
Ort	Municipality of the construction site	-
Projektname	Name of the building project	-
Datum	Date of the purchase	day.month.year
Bezeichnung	Identifier for the column on site	-
Stahlbetonstütze	Column type (see A.2)	-
Stahlbetonverbundstütze	Column type (see A.1)	-
rund	Cross section shape	-
eckig	Cross section shape	-

oval	Cross section shape	-
Durchmesser	Cross-sectional dimension	m
a	Cross-sectional dimension	m
b	Cross-sectional dimension	m
l	Column length	m
Nd	Load	kN
Md	Bending moment	kNm
Stück	Quantity	-
Hersteller	Manufacturer	-
intern	self-purchased	-
extern	externally purchased (subcontractor)	-
intern	Manufacturing cost	CHF/piece
Rabatt inkl. Skonto	Discounts	%
Transport	Transport costs	CHF/piece
intern inkl. Rabatt & Teuerung & Transport	Column costs	CHF/piece
extern	Delivery costs	CHF/piece
Rabatt inkl. Skonto	Discounts	%
extern inkl. Rabatt & Teuerung	Column costs	CHF/piece
A	Cross section area	m^2
V	Column volume	m^3
N/mm^2	Tension/Compressive stress	N/mm^2
CHF/m^3	Cost per volume	CHF/m^3
Preis Total inkl. Rabatt & Teuerung & Transport	Total column costs	CHF
Bemerkungen	Comments	-

Sign rule for load and bending moment: The column loads are mainly positive, which means the columns are subjected to a compressive force. Negative values mean the columns are subjected to a tensile force; this is very rare with concrete columns and could also indicate an input error.
Concrete columns are usually modeled as pin-ended columns and do not transmit bending moments. The impact of vehicles can cause bending moments in the column, which can be taken into account here in the table; the sign is irrelevant because the change of sign changes only the bending direction. The column must be able to absorb the bending moment in all directions.

5 Data Quality

The prediction accuracy should not fall below 70%. Otherwise, the construction costs for the offer would fluctuate too much. The share of column costs in an average building project is about 5% of the construction costs. With a prediction accuracy of 70%, there is a cost uncertainty of 1.5% on the total price.
It should be noted that due to the use of purchase prices, we are always slightly behind the market, and the most current market conditions are not included in the table.

The following procedure is planned to reach the desired accuracy:

1. clean up the data (remove obvious input errors)
2. model test with all data
3. split data into column types
4. split data into column types and cross-section shapes

If the desired accuracy cannot be achieved, it must be decided on a case-by-case basis whether the prediction accuracy is acceptable for the current construction project.

Example: If the accuracy of the prediction is 60% and we require accuracy in the final price of 1.5%, the share of the columns in the final price must not exceed 3.75%.

6 Data Flow

After the concrete columns have been awarded by the purchasing department, we get access to the order (PDF). After that, the values are transferred to the Excel file (manually); this is done for several projects together at a time when no project work is pending. The updated list can be read and cleaned with Python (Jupyter Notebook). The data can now be analyzed and used for modeling and plotting. If the prediction accuracy meets the desired needs, the updated model can be used to predict column prices.

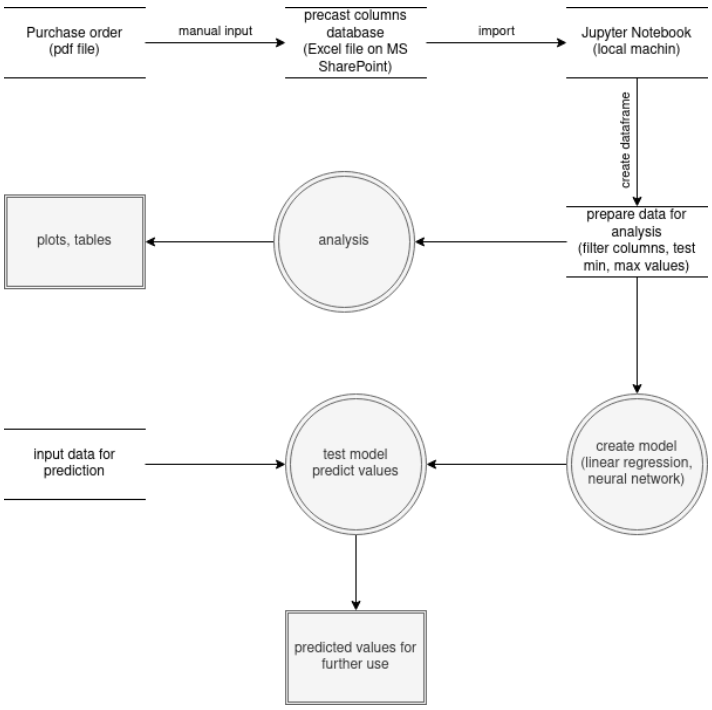

FIGURE 6.1: Data Flow

7 Data Model

Conceptual

To develop a tool for easy and reliable prediction of concrete column prices to simplify estimating construction projects.

Logical

Two models will be used:

1. linear regression
2. neural networks

The following data frames are intended:

TABLE 7.1: Intended data frames

variable name	**data frame**
raw_df	as read by pandas from the file
df	cleaned
rund	round columns
eckig	angular columns
oval	oval columns
sbv	steel/concrete composite columns
sb	reinforced concrete columns
sbv_rund	round steel/concrete composite columns
sbv_eckig	angular steel/concrete composite columns
sbv_oval	oval steel/concrete composite columns
sb_rund	round reinforced concrete columns
sb_eckig	reinforced concrete columns
sb_oval	reinforced concrete columns

It is planned to create the models for all the above data frames (expect raw_df) to determine which data frames give the best results.

The following columns are considered for the analysis, as they should have the most significant impact on price. If the desired accuracy cannot be achieved, other columns/features should be considered.

TABLE 7.2: Considered columns/feature

column/feature	**variable name**
load	Nd
cross-section area	A
column length	l
volume	V
unit price	Stückpreis

For the neural network, a normalization of the values is performed so that the magnitude of all values is in the same range (from 0 to 1)

TABLE 7.3: Normalization of columns/feature

column/feature	**adjustment**
load	Nd/10000
cross-section area	A x 10
column length	l/10
volume	V
unit price	Stückpreis/10000

Physical

There are no high demands on the hardware; only the neural network training requires an apparent computational effort.

8 Risks

Two main risks have been identified:

1. Data is entered into the excel file by hand. Here, there is a possibility of transcription errors. The incorrectly entered values can only be detected afterward if they deviate to the extent that they cannot correspond to reality.

2. There is the risk that the column prices depend primarily on the suppliers' economic situation/plant utilization, and the suppliers' profit is not always in the same frame. Thus the prices are not or only in subordinate frames dependent on physical sizes.

Input errors can be checked in Python per column (min/max values). The price margin of the supplier cannot be compensated and remains a risk.

Furthermore, the current inflation due to the Covid pandemic and the war in Ukraine cannot be taken into account. Here the global inflation surcharges of the manufacturers have to be factored in; this can be done in the source excel file.

9 Preliminary Studies

Descriptive statistic plots are shown below.
For the scatterplots, load and volume have been chosen as variables because the load is the critical structural variable, and volume is the product of the geometric variables. As there are many points in a small area, the alpha value of each point has been set to between 0.3 and 0.5

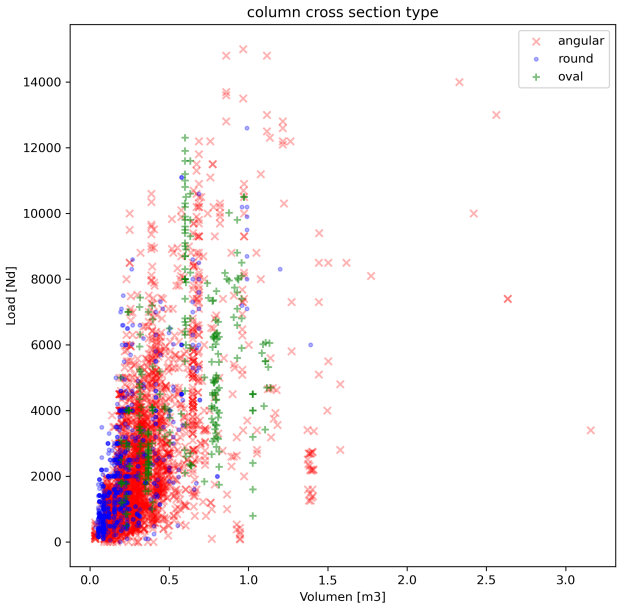

FIGURE 9.1: column cross-section types

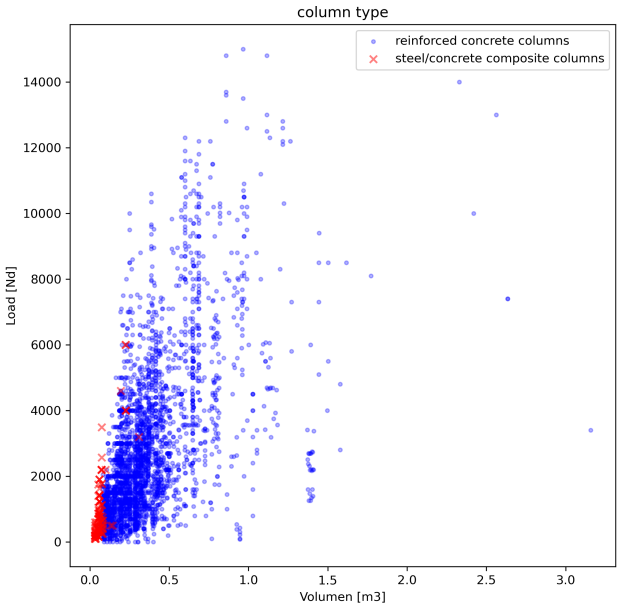

FIGURE 9.2: column types

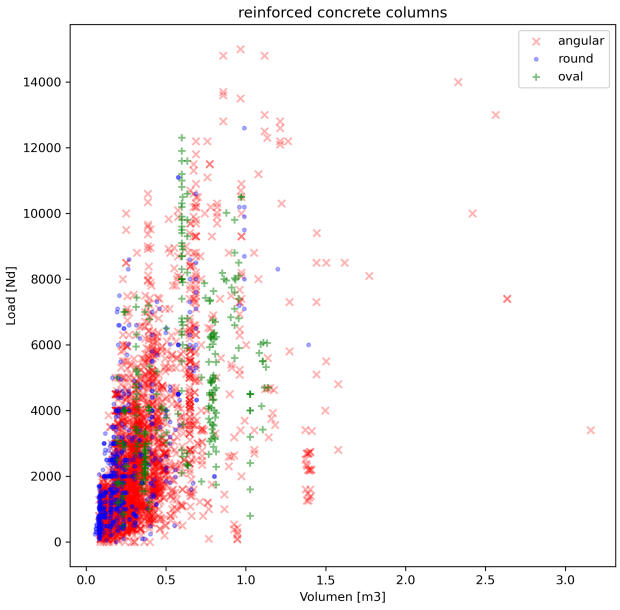

FIGURE 9.3: reinforced concrete columns by cross section type

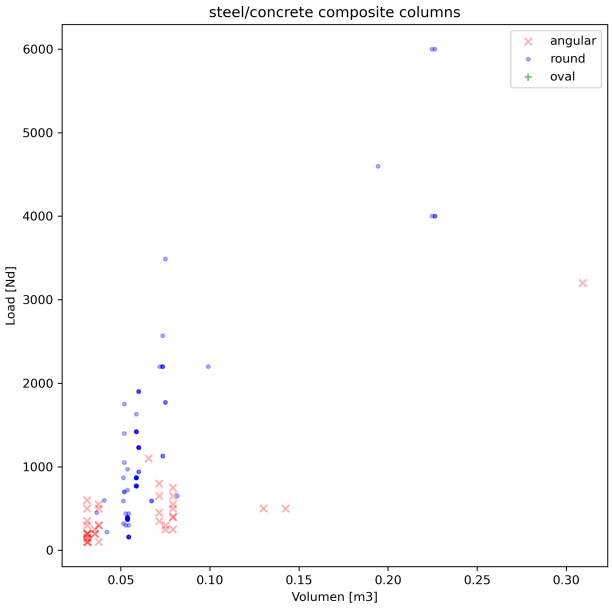

FIGURE 9.4: steel/concrete composite columns by cross section type

The first results indicate that depending on how the train/test split is done; the reliability will be achieved or not. Further investigation is needed.

```
x_train, x_test, y_train, y_test = cas.create_test_train_set(x, y)
```

```
LM = cas.LinearModel()
LMmodel = LM.linear_model(x_train, y_train)
print("MSE:")
print(f"Train: {LM.mse(x_train, y_train, LMmodel):0.0f}")
print(f"Test: {LM.mse(x_test, y_test, LMmodel):0.0f}")
print("\nR2:")
print(f"Train: {LM.r2(x_train, y_train, LMmodel):0.4f}")
print(f"Test: {LM.r2(x_test, y_test, LMmodel):0.4f}")

# LMmodel.intercept_
# LMmodel.coef_
```

```
MSE:
Train: 337
Test: 309

R2:
Train: 0.7125
Test: 0.6813
```

FIGURE 9.5: Linear Regression Model, first run including train/test split

```
MSE:
Train: 330
Test: 337

R2:
Train: 0.7055
Test: 0.7146
```

FIGURE 9.6: Linear Regression Model, second run including train/test split

```
MSE:
Train: 319
Test: 379

R2:
Train: 0.7094
Test: 0.6959
```

FIGURE 9.7: Linear Regression Model, third run including train/test split

10 Conclusion

The linear regression does not seem to achieve the desired accuracy at the moment; further refinements are to be made. The range of values of the input variables could be further restricted, which means a limitation in the application (not all ranges can be covered). This would be a feasible way and is an acceptable limitation of use.
Certain input data splits have a small number of columns (e.g., steel/concrete compound columns, 135 pieces), making prediction difficult/impossible.

The results of the neural network are still pending.

A Colum types

A.1 steel/concrete composite columns

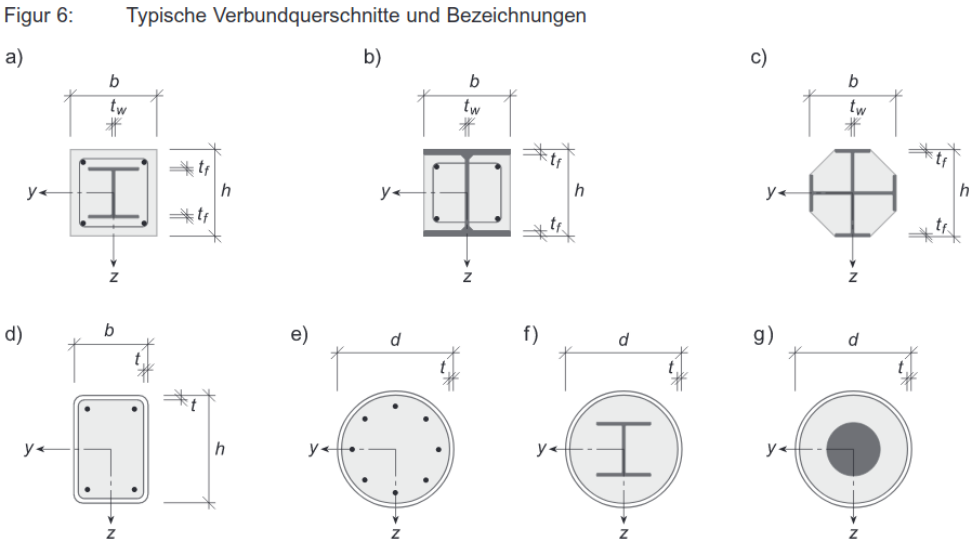

FIGURE A.1: SIA 264 [8] Figure 6, steel/concrete composite columns

A.2 reinforced concrete columns

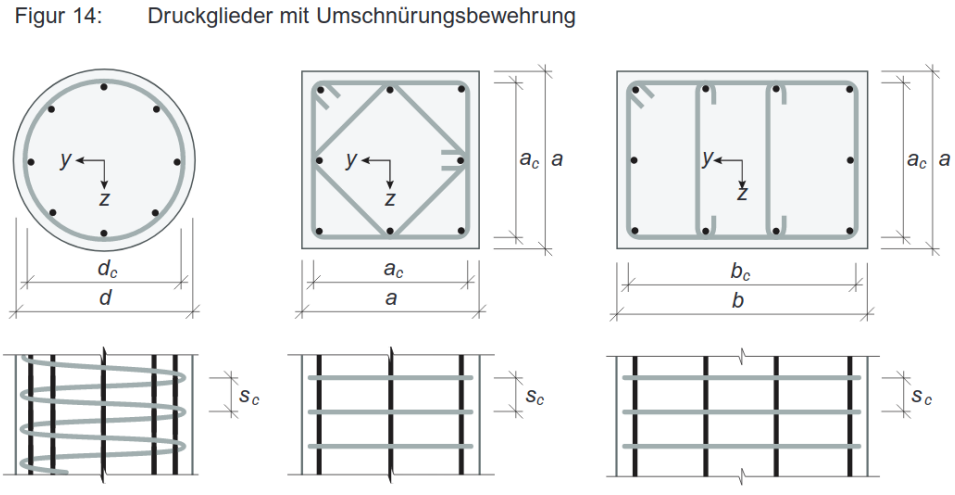

FIGURE A.2: SIA 262 [9] Figure 14, reinforced concrete columns

Reference

[1] *Jupyter*. URL: https://jupyter.org/ (visited on 10/28/2022).

[2] *Python*. URL: https://www.python.org/ (visited on 10/28/2022).

[3] *pandas - Python Data Analysis Library*. URL: https://pandas.pydata.org/ (visited on 10/28/2022).

[4] *NumPy*. URL: https://numpy.org/ (visited on 10/28/2022).

[5] *Matplotlib - Visualization with Pytho*. URL: https://matplotlib.org/ (visited on 10/28/2022).

[6] *scikit-learn: machine learning in Python*. URL: https://scikit-learn.org/stable/ (visited on 10/28/2022).

[7] *TensorFlow*. URL: https://www.tensorflow.org/ (visited on 10/28/2022).

[8] Normenkommission SIA 264. *SIA 264 Stahl-Beton-Verbundbau*. Schweizerischer Ingenieur- und Architektenverein, 2014.

[9] Normenkommission SIA 262. *SIA 262 Betonbau*. Schweizerischer Ingenieur-und Architektenverein, 2013.

List of Figures

List of Tables

List of Abbreviations

mm	Millimeter
m	Meter (1m = 1'000mm)
m²	Square meter
m³	Cubic meters
A	Area
V	Volume
N	Newton ($\frac{kg \cdot m}{s^2}$)
kN	Kilonewton (1kN = 1'000N)
kNm	Kilonwetonmeter
N/mm²	Newton per square millimeter
CHF	Schweizer Franken
SIA	Schweizerischer Ingenieur- und Architektenverein