Concrete compressive strength prediction with machine learning

Pedro Bernardino Alves Moreira

April 22, 2020

Abstract

Compressive strength is the main characteristic of concrete. The correct prediction of this parameter means cost and time reduction. This work built predictive models for 6 different ages of concrete samples (3, 7, 14, 28, 56, and 100 days). A set of data obtained in previous studies was used, a total of 1030 samples, with 9 variables: compressive strength, age, and 7 ingredients (water, cement, fine aggregate, coarse aggregate, fly ash, blast furnace slag, and superplasticizers). Another 6 variables were added to represent the proportions of the main ingredients in each sample (water/cement, fine aggregate/cement, coarse aggregate/cement, fine aggregate/coarse aggregate, water/coarse aggregate, and water/fine aggregate). The predictive models were developed in *R* language, using the *caret* package with the *Parallel Random Forest* algorithm and repeated cross-validation technique to optimize the parameters. The results were satisfactory and compatible with other studies using the same data set. The most important model, 28 days old, obtained *RMSE* of 4.717. The 3-day model obtained the best result, *RMSE* of 3.310. The worst result was the 56-day model, with *RMSE* of 5.939. The work showed that the compressive strength of concrete can be predicted. The choice of creating a model for each age, instead of using age as a predictor, allowed to get compatible results with the available data at each age. It was a promising alternative since good results were achieved by training with just one algorithm. This work facilitates exploration and new efforts to predict the compressive strength of concrete, it can be replicated using different algorithms or the combination of several.

Contents

List of Figures

List of Tables

1 Introduction

Compressive strength is the main characteristic of concrete, measured by tests of international standards that consist of the breaking of specimens. Measurement at 28 days is mandatory and represents the grade of the concrete. Knowing in advance what the result will be obtained for a given age, based on the proportions of its ingredients, is of great interest to concrete manufacturers, construction companies, and civil engineers.

This compressive strength is a nonlinear function of its ingredients and age, making it difficult to establish an analytical formula, although some formulas have already been proposed. (**???**) proposed a mathematical model to predict from the results of tests of 7 and 14 days, and (**???**) from 7 days. However, machine learning techniques can be used to model this characteristic from real sample data, using only the ingredients.

Many previous studies use the same dataset used by Yeh (n.d.) to predict the compressive strength of concrete. (**???**) got good results with the regularized extreme learning machine (RELM) technique, and (**???**) got even better results with the Artificial Neural Networks and cross-validation technique. This set of samples is so well known that there are many pages on the internet of unpublished studies that use it and have good results, such as (**???**), (**???**), (**???**) and (**???**). At the end of the work, the results found are compared to the works cited here.

Unlike previous studies with this dataset, this work does data preparation differently. The age of the concrete is the most unique feature that contributes to its compressive strength. For this reason, age is treated separately in the machine learning models, creating models for each age group.

2 Materials and Methods

2.1 Materials

The methodology was carried out using RStudio software (**???**), an integrated virtual environment for code development in *R* (**???**). Throughout the process, all code executed was documented in the same order as its execution in Appendix 2, and reference was always made to codes throughout the text. All relevant information related to the operating system and installed packages has been presented in Appendix 1.

2.2 Reproducibility

In order to guarantee reproducibility, whenever there was some code that could use probabilistic operations, a *seed* was defined before its execution, ensuring that when run on another machine, with the same version of *R* and the same *seed*, get the same result. The *seeds* can be checked throughout Appendix 2.

2.3 Obtaining the data

The data was downloaded from the University of California Irvine website (**???**) (**??**). In total there are 1030 samples with 9 columns. The samples were renamed and an id column was added to facilitate data manipulation (**??**). The columns were reordered to put the new id column in the first position (**??**). The first samples can be viewed in the table 1.

Table 1: First 6 samples

ID	Cement kg/m^3	B.F.S. kg/m^3	Fly ash kg/m^3	Water kg/m^3	Superp. kg/m^3	C.Aggregate kg/m^3	F.Aggregate kg/m^3	Day	Comp.Str. MPa
1	540.0	0.0	0	162	2.5	1040.0	676.0	28	79.99
2	540.0	0.0	0	162	2.5	1055.0	676.0	28	61.89
3	332.5	142.5	0	228	0.0	932.0	594.0	270	40.27
4	332.5	142.5	0	228	0.0	932.0	594.0	365	41.05
5	198.6	132.4	0	192	0.0	978.4	825.5	360	44.30
6	266.0	114.0	0	228	0.0	932.0	670.0	90	47.03

2.4 Data preparation

The preparation of the data consisted of transforming the sample set in order to maintain only relevant data for the subsequent studies. Data that were considered irrelevant or that had the potential to add undesirable noise to the analysis were removed. In addition, the relevant data has been transformed to better fit the studies in the next steps.

2.4.1 Initial data cleaning

Initially, there were 25 duplicate samples that were removed, resulting in a new total of 1005 samples (**??**).

The data show the variables in the columns and samples in the rows. However it was found that some samples are identical in proportions of ingredients, changing only the value of age and compressive strength, for example, samples 653, 654, 678 and 681, shown in the table 2.

Table 2: Samples with same composition

ID	Cement kg/m^3	B.F.S. kg/m^3	Fly ash kg/m^3	Water kg/m^3	Superp. kg/m^3	C.Aggregate kg/m^3	F.Aggregate kg/m^3	Day	Comp.Str. MPa
653	102	153	0	192	0	887	942	3	4.57
678	102	153	0	192	0	887	942	7	7.68
681	102	153	0	192	0	887	942	28	17.28
654	102	153	0	192	0	887	942	90	25.46

In addition, there are also samples with the same values and proportions of ingredients, but with different compressive strength, probably due to differences in the building process. This is the case, for example, of samples 472, 473 and 474, shown in the table 3.

To facilitate the analysis of the samples, all samples that are the same in relation to the ingredients, have been assigned the same *id*. In addition, as the compressive strength at 28 days is the parameter to determine the grade of the concrete, only the elements containing that day among its samples were maintained. In the case of the same samples but with different results of compressive strength, the values were averaged. After all these changes (**??**), the new total of samples was reduced to 970, containing 416 different settings for the proportions of ingredients.

Table 3: Same samples with different results

ID	Cement kg/m^3	B.F.S. kg/m^3	Fly ash kg/m^3	Water kg/m^3	Superp. kg/m^3	C.Aggregate kg/m^3	F.Aggregate kg/m^3	Day	Comp.Str. MPa
472	446	24	79	162	11.6	967	712	28	57.03
473	446	24	79	162	11.6	967	712	28	44.42
474	446	24	79	162	11.6	967	712	28	51.02

The result can be seen in the table 4. All samples with equal ingredient settings have the same id, and when they had different results for the same days, they were transformed into just one sample, with the arithmetic mean in the compressive strength.

Table 4: Previous samples after processing

ID	Cement kg/m^3	B.F.S. kg/m^3	Fly ash kg/m^3	Water kg/m^3	Superp. kg/m^3	C.Aggregate kg/m^3	F.Aggregate kg/m^3	Day	Comp.Str. MPa
472	446	24	79	162	11.6	967	712	28	50.82
653	102	153	0	192	0.0	887	942	3	4.57
653	102	153	0	192	0.0	887	942	7	7.68
653	102	153	0	192	0.0	887	942	28	17.28
653	102	153	0	192	0.0	887	942	90	25.46

2.4.2 Age selection

As previously described, the main age for analysis of compressive strength is 28 days, but other ages can also be used to build predictive models. However, it is necessary to verify how relevant the data of these other ages are. Starting with the distribution of the samples in relation to each age (**??**) shown in the figure 1.

It was observed that the ages of 90, 91 and 100 days probably represent extremes to each other in the ingredient configurations, since they are relatively close ages but with very different values, especially for 90 and 91.

This hypothesis was verified using the principal component analysis method, applied to samples of these 3 ages (**??**). The figure 2 shows how the samples relate to each other (which are similar or different) and revealed how each variable contributes to the analysis. The first two dimensions represent 37% and 24% $ respectively of the variance.

Another important point considered, of the concrete nature itself, is the fact that the growth rate of its compressive strength decreases with time, reaching a certain stability value. The figure 3 shows the compressive strength over the days for samples with more than 5 data, that is, data available for at least 6 different ages (**??**).

For the reasons presented in the figures 1, 2 and 3, it was considered that the ages of 90, 91 and 100 days can be grouped to improve reading and decrease sample noise. They were converted to the same value, the age of 100 days was chosen (**??**). As shown in the figure 3, the resistance only increases, after 100 days the resistance to compression will be greater than or equal to the value of 90 or 91 days.

Figure 1: Boxplot - Compressive strength (MPa) vs age (days)

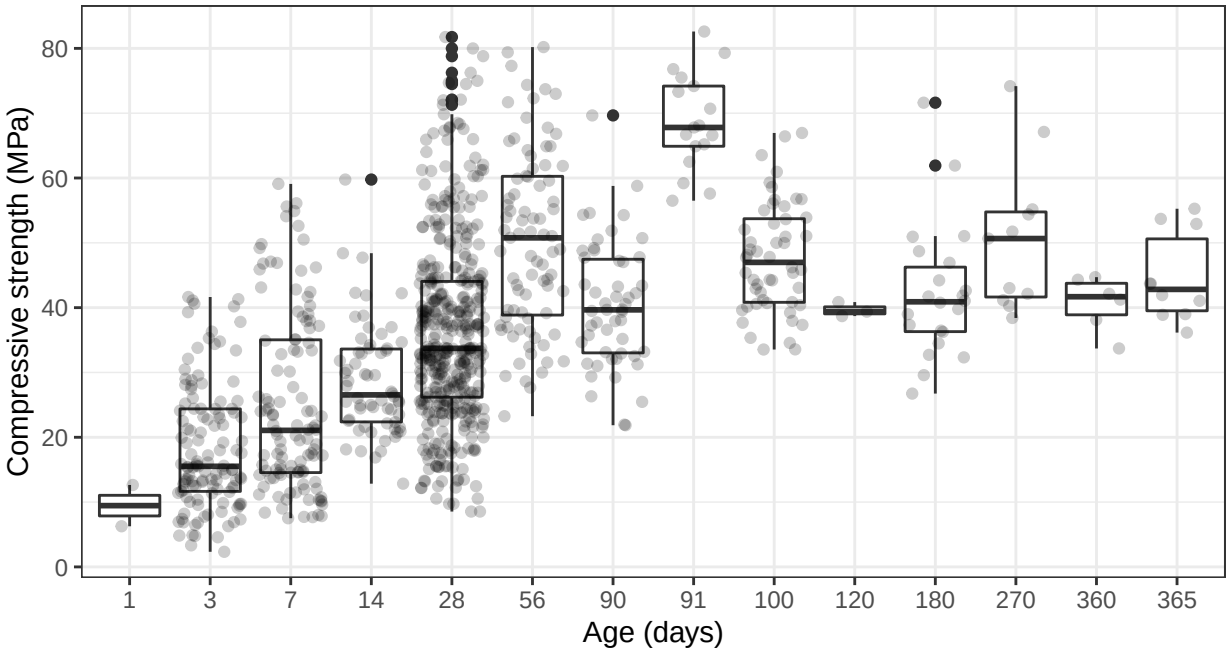


Figure 2: Principal component analysis - 90, 91 e 100 days

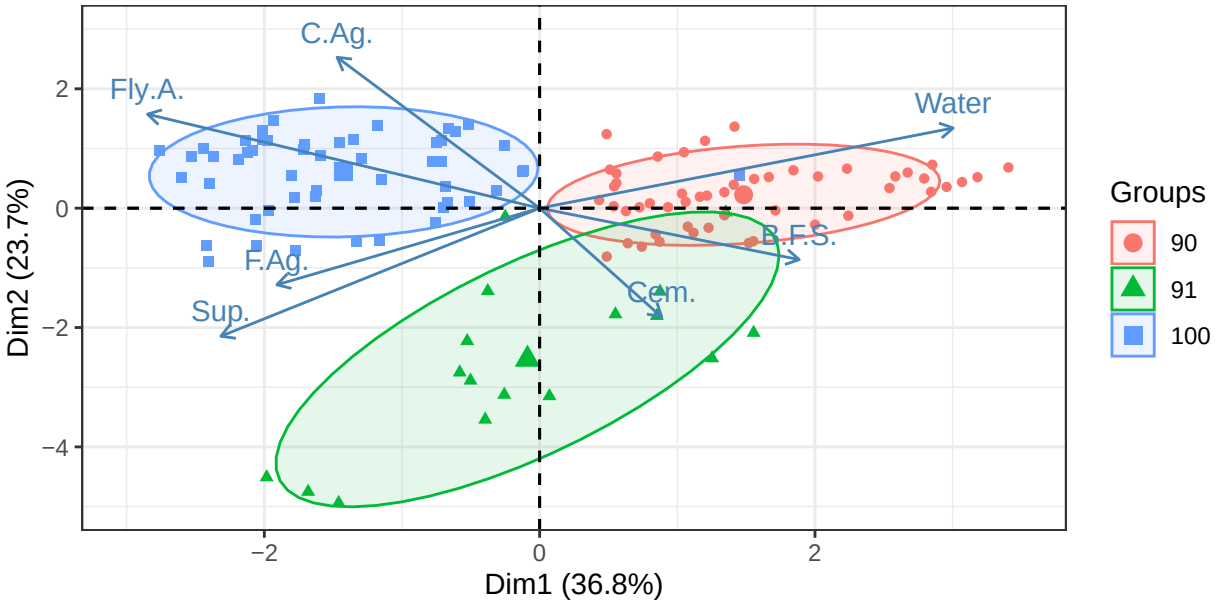
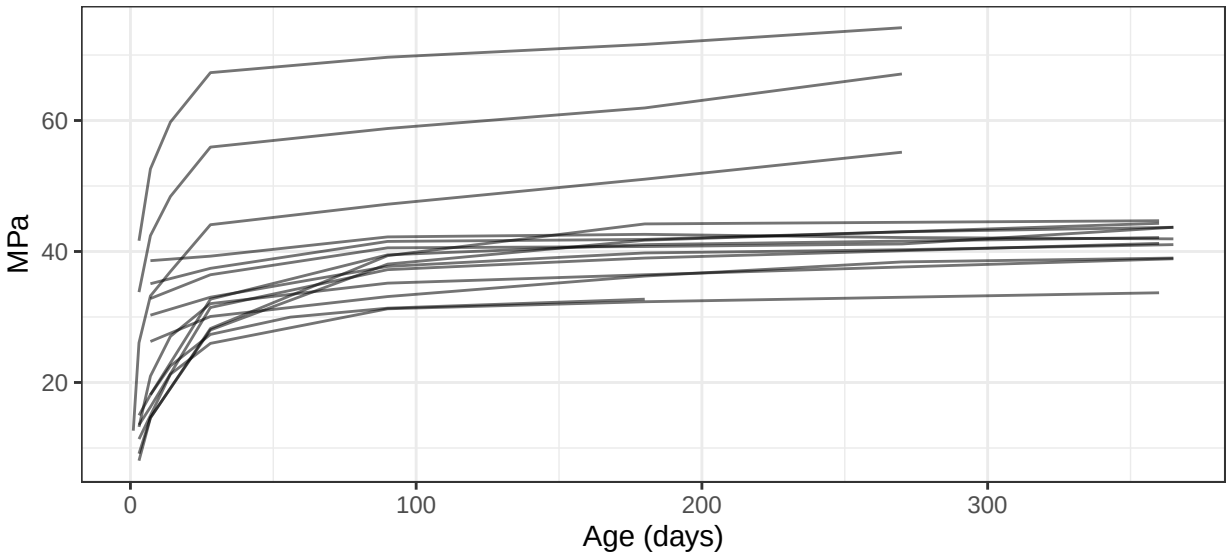
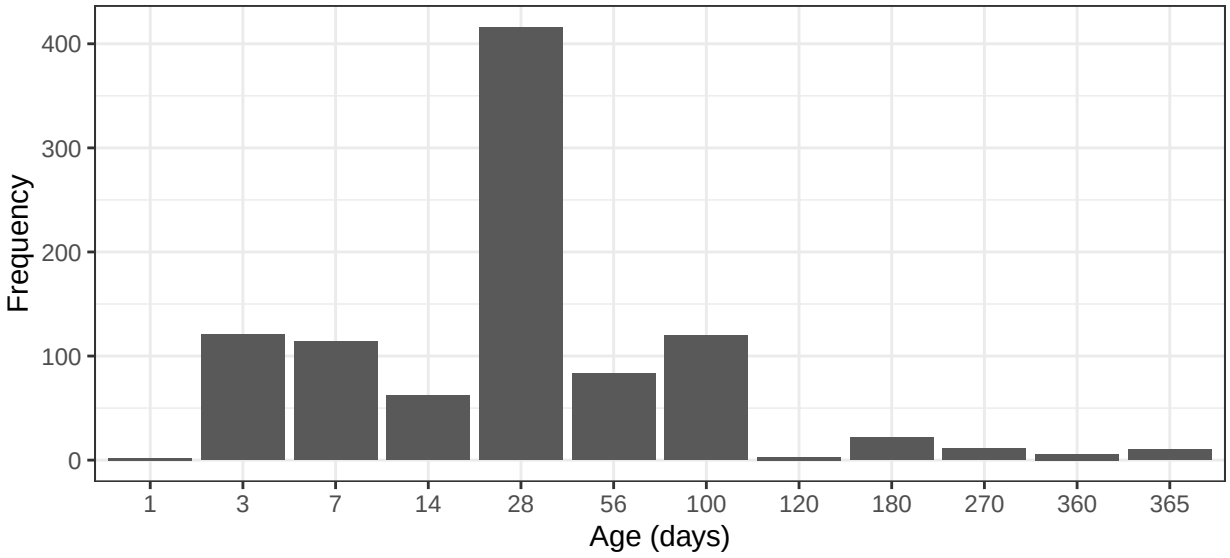


Figure 3: Compressive strength through time



Another topic analyzed in the selection of ages was the observed frequency of each age value after this transformation from 90, 91 in 100 days, shown in the figure 4. Some values of days have very low concentrations of samples, at the risk of damaging more than helping to create the models, so they were removed (**??**). The criterion adopted was to maintain only ages with a frequency greater than 50, only the values of 3, 7, 14, 28, 56 and 100 days.

Figure 4: Ages frequency



2.4.3 Data reorganization

The samples were grouped to maintain only one sample from each set of configuration of the proportions of the ingredients, adding new variables/columns for resistance at each age (**??**). The result in the first

samples after this processing is shown in the table 5.

Table 5: First 6 samples after reorganization

ID	Cement kg/m^3	B.F.S kg/m^3	Fly ash kg/m^3	Water kg/m^3	Superp. kg/m^3	Coarse Ag. kg/m^3	Fine Ag. kg/m^3	3 days MPa	7 days MPa	14 days MPa	28 days MPa	56 days MPa	100 days MPa
1	540.0	0.0	0	162	2.5	1040.0	676.0				79.99		
2	540.0	0.0	0	162	2.5	1055.0	676.0				61.89		
3	332.5	142.5	0	228	0.0	932.0	594.0		30.28		33.02		37.72
5	198.6	132.4	0	192	0.0	978.4	825.5	9.13	14.64		28.02		38.07
6	266.0	114.0	0	228	0.0	932.0	670.0				45.85		47.03
7	380.0	95.0	0	228	0.0	932.0	594.0		32.82		36.45		40.56

The number of samples and distinct samples after all this manipulation remained the same, a total of 416 (**??**).

2.4.4 Adding new variables

To finish the data preparation, new columns were added to the dataset (**??**). Starting with the concrete class, for example if the compressive strength is between 25 and 30, it receives the class *C25*. The inclusion of the class was important because the compressive strength in *MPa* is a continuous variable, which will be used in the regression models, but the class as a discrete variable can provide another visualization of the data. The approximate mix of concrete was also added, which represents the proportions of aggregates (fine and coarse) for cement. Other proportions between the main ingredients were also added. The new variables are presented in the table 6.

Table 6: New features

ID	Class	Approximated Mix	Water / Cement	Fine Ag. / Cement	Coarse Ag. / Cement	Fine Ag. / Coarse Ag.	Water / Coarse Ag.	Water / Fine Ag.
1	C75	1:1:2	0.3000	1.2519	1.9259	0.6500	0.1558	0.2396
2	C60	1:1:2	0.3000	1.2519	1.9537	0.6408	0.1536	0.2396
3	C30	1:2:3	0.6857	1.7865	2.8030	0.6373	0.2446	0.3838
5	C25	1:4:5	0.9668	4.1566	4.9265	0.8437	0.1962	0.2326
6	C45	1:3:4	0.8571	2.5188	3.5038	0.7189	0.2446	0.3403
7	C35	1:2:2	0.6000	1.5632	2.4526	0.6373	0.2446	0.3838

2.5 Data visualization

In order to assess the need for further manipulation before building the models, in this step the 416 samples already processed were visualized and analyzed.

2.5.1 Descriptive statistics

The table 7 presents the statistical data of the continuous variables (**??**). The *Null* line represents the number of zeroed values for the ingredients, and the *NA* line represents the number of missing data. As the samples were filtered to maintain only sets of samples with known values of compressive strength at 28 days, the number of *NAs* is zero for that age. The figure 5 presents the statistical data of the discrete variables (**??**).

Table 7: Descriptive statistics - continuous variables

	Samples	Null	NA	Min	Max	Range	Sum	Median	Mean	SE mean	CI mean	Variance	Std.Dev.	Coef.Var
Cement	416	0	0	102.00	540.00	438.00	109373.10	257.70	262.92	5.10	10.02	10817.50	104.01	0.40
B.F.S.	416	174	0	0.00	359.40	359.40	35824.60	94.25	86.12	4.32	8.49	7755.00	88.06	1.02
Fly ash	416	202	0	0.00	200.10	200.10	26389.00	71.25	63.44	3.26	6.40	4407.81	66.39	1.05
Water	416	0	0	121.80	247.00	125.20	76335.60	185.00	183.50	0.94	1.86	370.73	19.25	0.10
Superplast.	416	107	0	0.00	32.20	32.20	2871.30	7.60	6.90	0.26	0.52	28.85	5.37	0.78
Coarse agg.	416	0	0	801.00	1145.00	344.00	397799.90	953.35	956.25	4.12	8.10	7063.06	84.04	0.09
Fine agg.	416	0	0	594.00	992.60	398.60	317809.80	769.65	763.97	3.59	7.06	5371.89	73.29	0.10
3 days	121	0	295	2.33	41.64	39.31	2210.82	15.52	18.27	0.87	1.72	91.64	9.57	0.52
7 days	114	0	302	7.51	59.09	51.58	2845.52	21.06	24.96	1.29	2.55	188.81	13.74	0.55
14 days	62	0	354	12.84	59.76	46.92	1782.56	26.54	28.75	1.10	2.19	74.62	8.64	0.30
28 days	416	0	0	8.54	81.75	73.21	15101.13	33.72	36.30	0.70	1.38	206.30	14.36	0.40
56 days	83	0	333	23.25	80.20	56.95	4178.77	50.77	50.35	1.52	3.02	190.82	13.81	0.27
100 days	120	0	296	21.86	82.60	60.74	5701.90	45.61	47.52	1.17	2.31	163.40	12.78	0.27
Water / Cement	416	0	0	0.27	1.88	1.62	340.60	0.73	0.82	0.02	0.03	0.11	0.34	0.41
Fine agg. / Cement	416	0	0	1.14	9.24	8.10	1415.82	2.94	3.40	0.07	0.14	1.96	1.40	0.41
Coarse agg. / Cement	416	0	0	1.55	8.70	7.14	1761.06	3.67	4.23	0.08	0.16	2.70	1.64	0.39
Fine agg. / Coarse agg.	416	0	0	0.53	1.16	0.63	335.42	0.80	0.81	0.01	0.01	0.01	0.11	0.14
Water / Coarse agg.	416	0	0	0.12	0.29	0.17	80.66	0.19	0.19	0.00	0.00	0.00	0.03	0.16
Water / Fine agg.	416	0	0	0.13	0.38	0.26	101.26	0.24	0.24	0.00	0.00	0.00	0.04	0.17

2.5.2 Correlation between ingredients and compressive strength

The figure 6 shows the correlation of variables for each set of ages (**??**). The figure 7 presents the same data, but instead of correlating them all, it only correlates with the compressive strength, showing the values in more detail (**??**).

The interpretation of the figure 7 suggests that the strength of the concrete is positively related mainly to the cement and superplasticizer ingredients and negatively to the water and fine aggregate. The smaller the amount of cement for aggregates and water, the more negatively they are correlated with compressive strength.

The figures 8 and 9 show the relationship between the main ingredients (known as mix) in relation to the compressive strength (**??** and **??**). The interpretation of these figures shows that the greater the amount of cement in relation to the other ingredients, the greater the resistance to compression.

2.5.3 Variables distribution

The figure10 shows the distribution of variables in the samples (**??**). It was calculated using data only at 28 days.

The figure 11 shows the distribution of ingredients and compressive strength for each set of ages (**??**), in case of the 28 days it presents the same information as the figure 10. It shows that the resistance to compression gradually increases over time, as expected. Furthermore it is seen that the concentration of the ingredients can vary a lot when stratified by ages.

Figure 5: Descriptive statistics - categorical variables

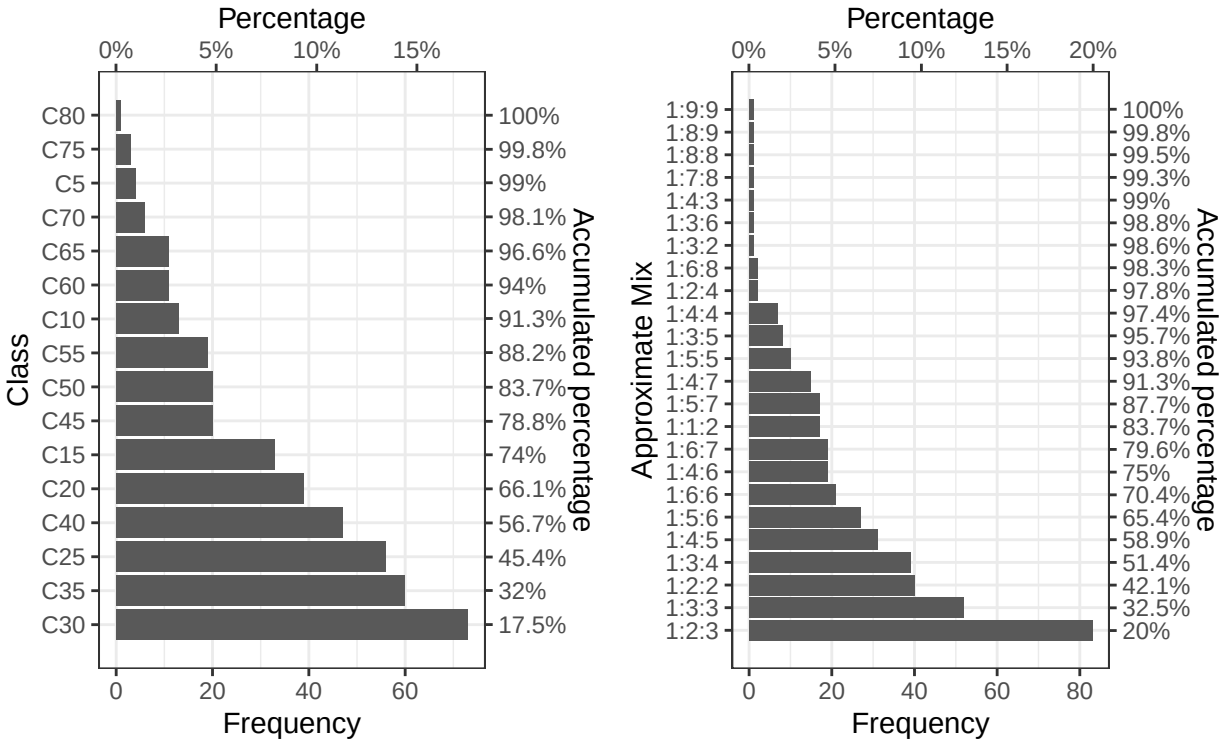


Figure 6: Correlations at each age

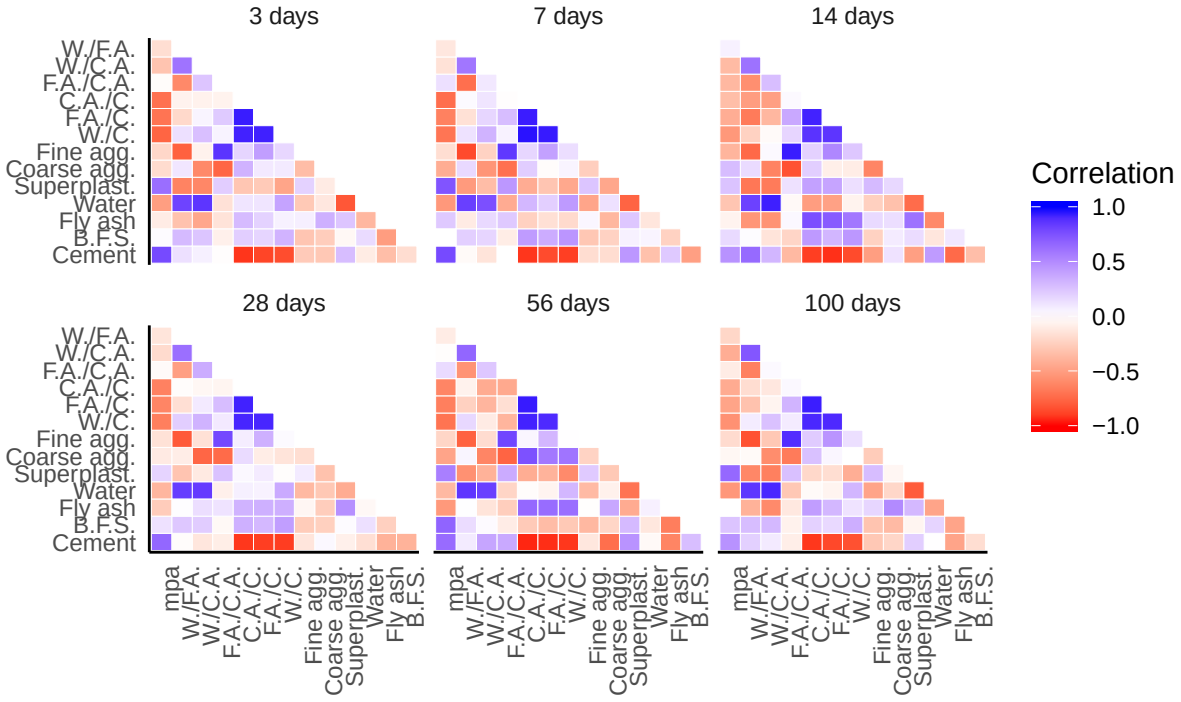


Figure 7: Correlation of variables with compressive strength over time

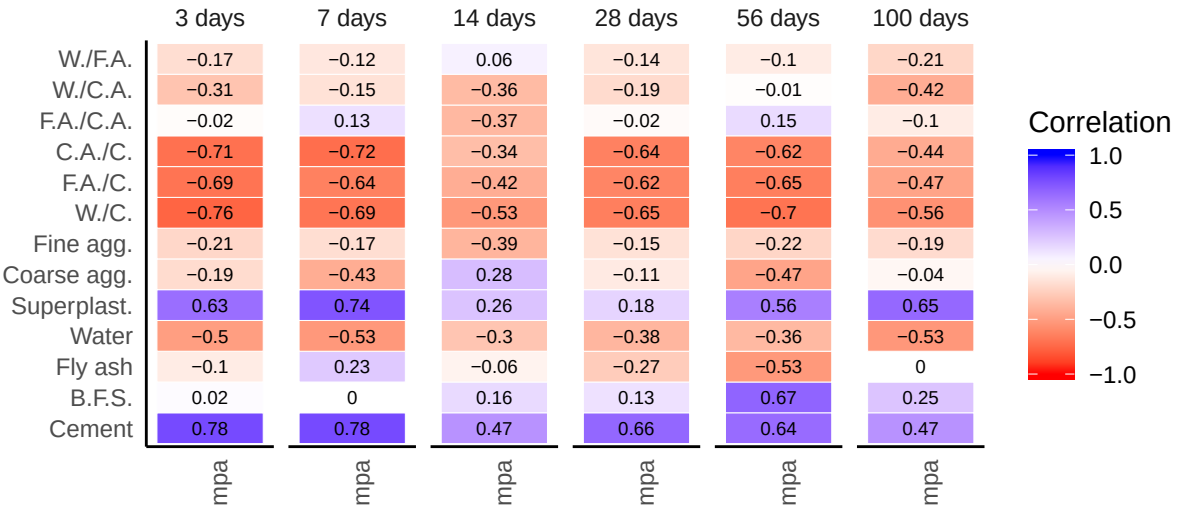


Figure 8: Relationship between approximated mix, water, MPa and age

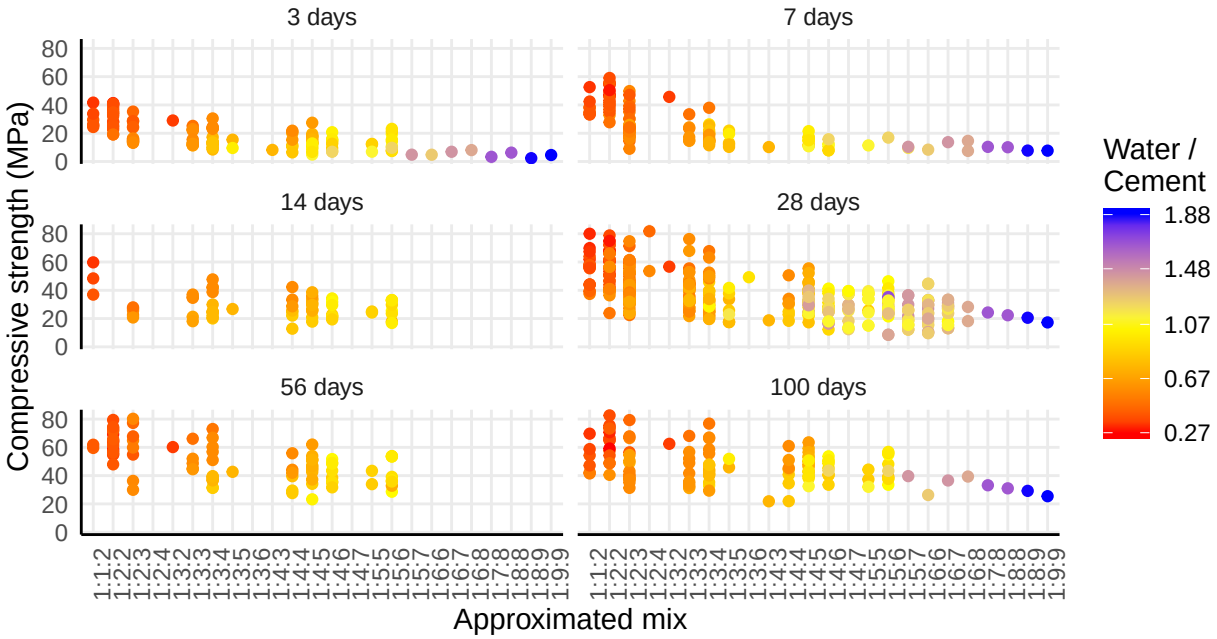
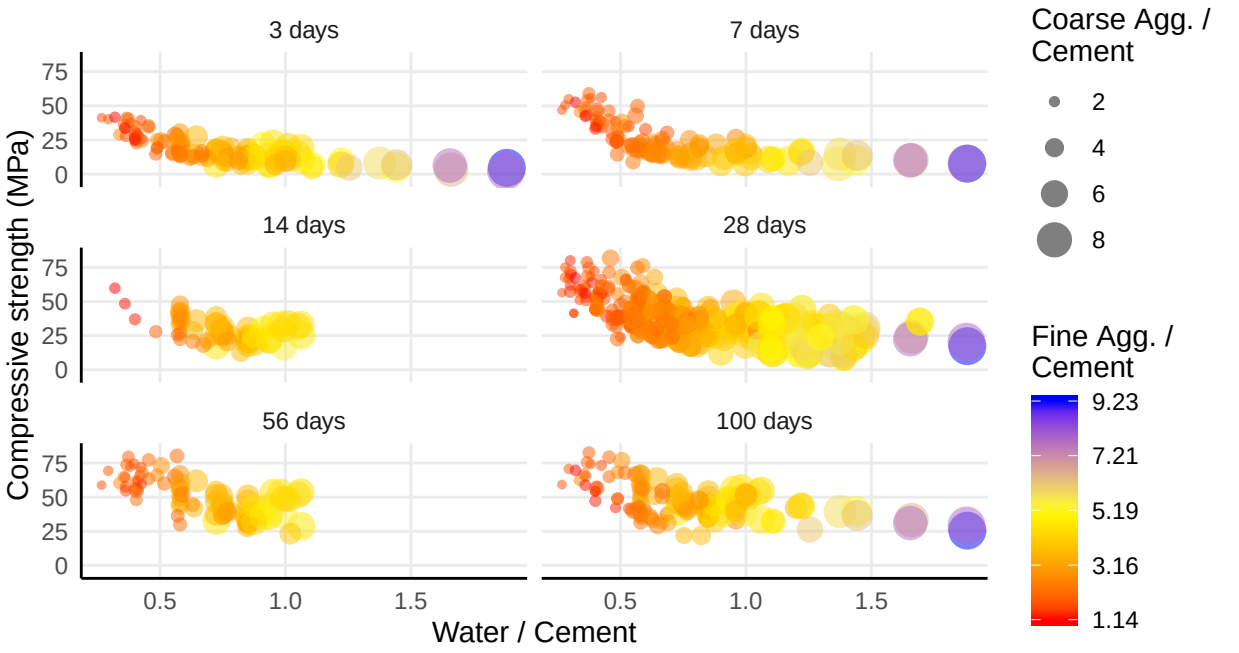


Figure 9: Relationship between concrete main features



2.5.4 Principal component analysis

In the figure 12, using an alternative classification, the principal component analysis was performed on the ingredients (**??**). The classification separates concrete into 4 different compressive strength groups, low up to *20 MPa*, normal up to *40 MPa*, medium up to *70 MPa* and high above that. It is possible to notice that the groups overlap, but there is a differentiation between the high and low group.

3 Literature cited

Yeh, I-Cheng. n.d. "Modeling of Strength of High-Performance Concrete Using Artificial Neural Networks." https://doi.org/10.1016/S0008-8846(98)00165-3.

Figure 10: Variables distribution

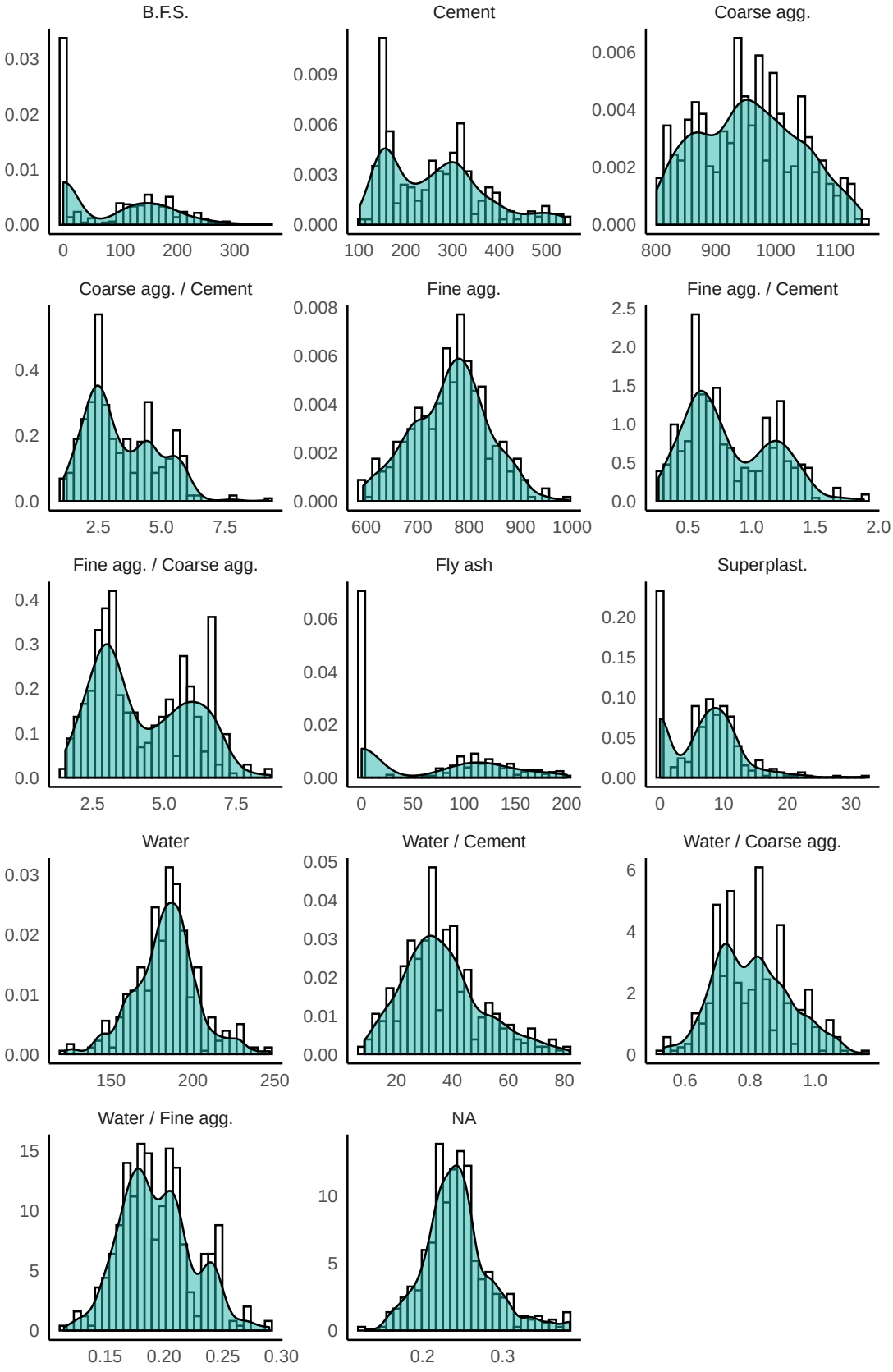


Figure 11: Variables distribution grouped by age

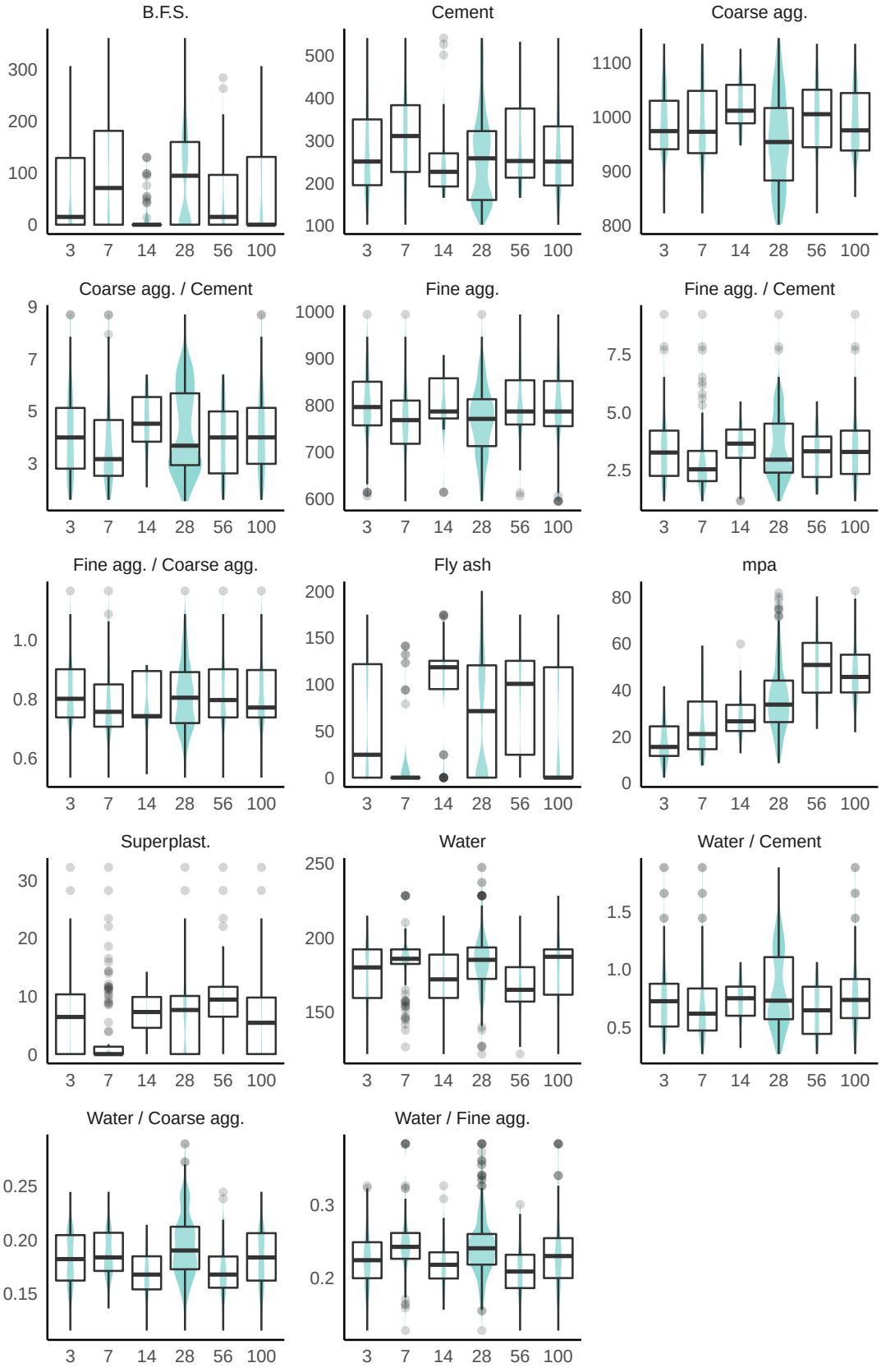


Figure 12: Principal component analysis on ingredients