LC3 Compressive Strength Analysis

David Alonso del Barrio, Francisco Javier Blázquez Martínez, Andrés Montero Ranc
Franco Zunino Sommariva
Construction Materials Laboratory, EPFL, Switzerland

***Abstract*—Cement industry is responsible of around 6% of CO2 emissions in the whole planet. LC3 stands for Limestone Calcined Clay Cement, a new type of cement which can reduce the emissions in it is elaboration by up to 40%. In this paper we are analyzing the compressive strength of this material depending on the properties of the clays involved in its preparation and finding the determining factors in LC3 composition. We provide models for estimating the compressive strength and its reliability at different stages concluding that it is a solid alternative and even a improvement to the classical cement.**

I. Introduction

This project is encompassed in the *Machine Learning* master course at the *École polytechnique fédérale de Lausanne*. In it, we provided data analysis tools to the Construction Materials Laboratory.

In this document we expose the process, ideas and decisions taken during the project development.

II. Data

At the beginning of the project we were given two excel files. These were not fully structured or organized with a given rigid format so, of course, after the first sight analysis, structuring the data was our first task.

On the one hand we had measurements of about 20 different features for 55 different types of clay from all over the world. These features included such disparate things as statistics of the particle size distribution, particle average surface, content of several chemical compounds, content of certain minerals... Unfortunately this dataset was not complete but some features had missing values in most of their entries.

On the other hand we had the compressive strength measurements for these clays after 1, 3, 7, 28 and 90 days from its preparation. We received only one pair compressive strenght-standard deviation for each clay and day, what implies that our original dataset consisted on less than sixty points for each of the measured days. These points proceeded from three different experiments reduced to a single pair average-standard deviation.

After asking the laboratory for the original and full data, they provided us with a series of excel files not intuitive at all, rather messy, with different units of measure and prepared by a person who was no longer in the laboratory. All this together made that data inaccessible at first.

III. Work Structure

Having these two datasets, we created two lines of work, the first one with the the averages and the standard deviation of the experiments (while preparing the full dataset) and the second one with all the measurements that we had available.

Also before starting the project, we realized that the lack of data was evident. We were dealing with up to 20 possible features for less than 150 points. After reading about how to handle this situation we decided follow these guidelines:

- Restrict to simple models
- Use feature selection
- Ensure data integrity and cleanliness
- Provide confidence intervals

IV. Data preprocessing

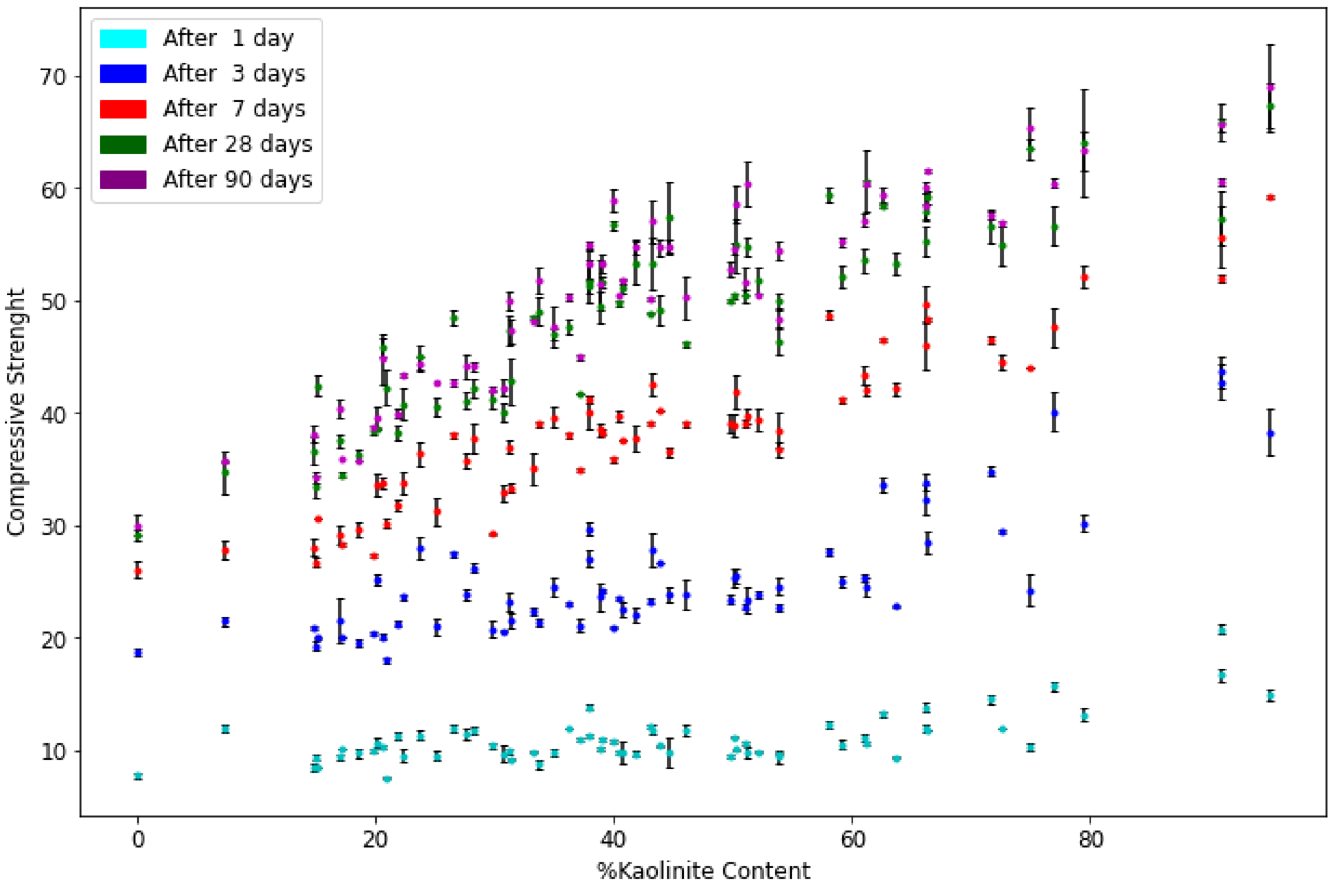
Here we include everything from having the unstructured data in the excel provided by the laboratory until a first data analysis.

First we started by preparing our data to be readable by a computer. We had to create IDs for the clays (which were appearing with different names), search and gather the measurements of the compression strength for each of these and for each day, deal with errors detected, deal with inconsistencies between both datasets... This is one of the parts that took us the longest because, even using code developed for automatically detecting errors and checking inconsistencies, it required a lot of human work and it would have been impossible without the help of an expert.

Once we had our data we visualized it and run some first analysis of correlations and outliers:

A. Correlations

This analysis was even more important with our shortage of data because it not only helped us to have an idea of those features with a high predictive value, but also to prevent adding features correlated to a model which could easily lead to overfitting in this case. We detected the *Kaolinite Content* as a feature with a high predictive value for every day measured (see **??**) as well as we detected other features highly correlated as $D90-SO_3$, $D90-Dv50$, $D10-MnO$. We will not go here into a deeper explanation of each feature.



B. Outliers

After visualizing the data, we appreciated a few possible outliers in the measurements at day 1 and 3. However, as all the experiments were repeated (in the laboratory at the EPFL, a reliable environment) and the behaviour was persistent we decided not to remove them. We did not remove either deviated pessimistic points since we are creating a model for construction materials and we considered more responsible to put ourselves in the worst case.

V. Kaolinite-based models

Once detected the high correlation of the *Kaolinite Content* with the compression strength for all the days contemplated. We decided to start creating simple models involving only this feature.

A. R-squared and MSE

R-squared is an almost perfect metric for knowing how good is our least squares model in this case. However, this metric is always improving as we add or create more variables to our regression so, for avoiding overfitting, we measure also the MSE of the model with leave one out cross validation which helps us to precisely estimate it with so little data.

B. Linear Regression

With this first approach we did not obtain good metrics for our models except for the day 7 model. This is because even when we know that there is a high correlation between *Kaolinite Content* and compression strength, this relation has not to be linear. Visualizing the data it is reasonable clear that we need a model more expressive to better fit the data.

C. Non Linear Models

Following that logic, we used feature augmentation to add *Kaolinite Content Square* to our model so we can fit better the distribution of the data. This notably improved our models (specially for the days 28 and 90). We did not continue the feature augmentation (adding the cubic term) because visualizing the data it was clear that it followed a distribution increasing the compression strength when increasing the *Kaolinite Content*. We wanted our model function to be increasing and considered not following having this behaviour a signal of overfitting.

It is precisely what happened in the models for days 1 and 3 where the parable vertex was inside of the range of the *Kaolinite Content*. However, we could see that it was caused by the irregular data distribution in the feature domain.

VI. Confidence analysis

The little amount of data makes this even more important and, this part is also the one that justifies the effort for obtaining a full dataset (more points is what has given us smaller confidence intervals). We have created lower bounds of certain probabilities for the compressive strength of the LC3 depending on the *Kaolinite Content* (graphics in the right, days 1 to 90 from up to down).

These have been obtained with the python library *statsmodels* and are confidence intervals for the linear regression parameters what, together with the fact that our points are not evenly distributed through all the *Kaolinite Content* domain, could lead to certain parts of the model not having an accurate bound locally. We have not however appreciated this.

VII. Feature selection

After creating the models based on the *Kaolinite Content* we wanted to take advantage of these (and their shape, close to the data distribution) but reducing the points sparsification or distance to the model so we decided to add more features. Once again, we could not add many of them because it would lead to overfitting. We created a function to decide which features complemented best the Kaolinite-based model.

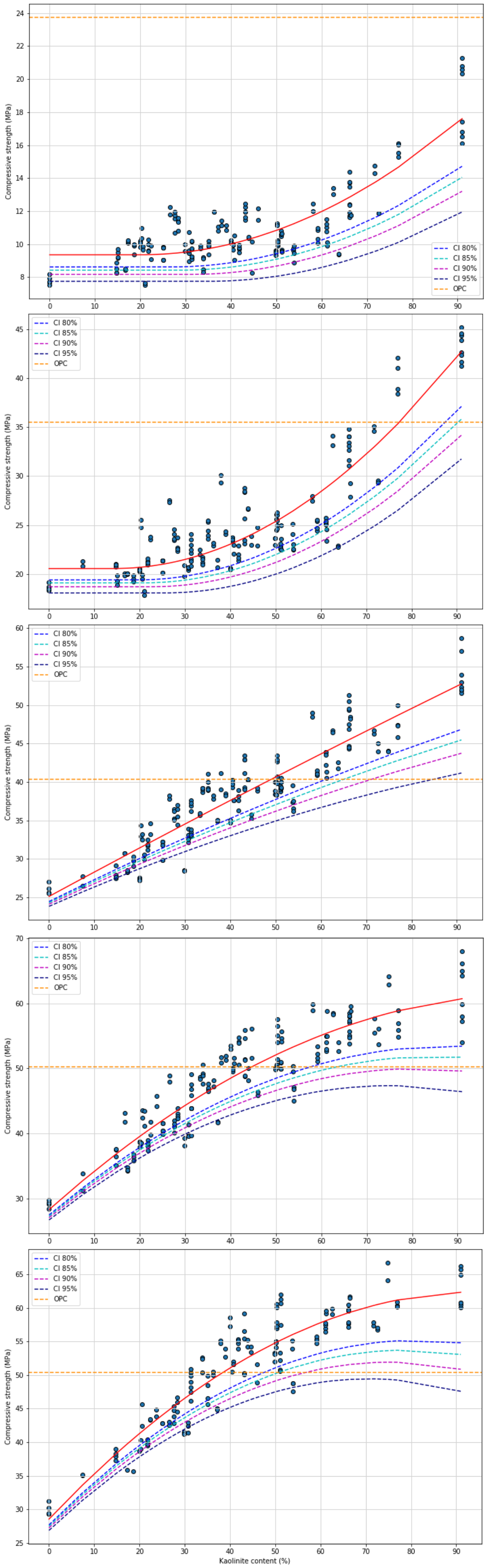
A. Adjusted R-squared

For deciding which was the best feature to be added to our model we used adjusted R-squared. It is a version R-squared adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than what would be expected by chance. It can decrease otherwise. Since in our case we only had a few points, it provided us with a way to penalize equations that take into account many variables, helping us to avoid overfitting. As always, we also considered MSE computed with leave one out cross validation for avoiding overfitting.

B. Relevant features

Since most of the features had missing values and this translated into dropped points, we realized that some features were having a bigger adjusted R-squared because after dropping missing values they had few points remaining (and of course it is easier to fit better less points with the same number of variables). We could not trust these features but it let us see their potential. We set a threshold for relying the features and got that the most relevant features were (ordered by relevance and reliability): *BET_specific_surface, span, D90, D10*

That is, the ones related with the size and shape of the particles in the clay are more relevant than those of the chemical composition (leaving aside the *Kaolinite Content*).



VIII. Multiparameter models

Continuing with this approach, and taking into account the reliable features, we improved the kaolinite-based models by adding these features.

	Intercept	Kaolinite	Kaolinite^2	D10	D90	SPAN	BET	adjusted R2
Day 1	9,9983	-0,0649	0,0016					0,7051
	0,1622	-0,3757	0,9814	-0,0564				0,8080
	0,1324	-0,3595	0,9952		0,0369			0,8040
	0,1494	-0,3938	0,9675			0,0573		0,7052
	0,1445	-0,2789	0,8748				-0,0063	0,7300
Day 3	21,3410	-0,1077	0,0038					0,8159
	0,1221	-0,3282	1,1413	-0,0881				0,8900
	0,0646	-0,3333	1,2087		0,1321			0,8970
	0,1150	-0,5741	1,3381			0,1874		0,7052
	0,1169	-0,2869	1,0961				-0,0236	0,7052
Day 7	25,1080	0,3188	-0,0002					0,8566
	0,0066	0,8585	-0,0326	0,0106				0,8930
	-0,0106	0,8371	0,0121		0,0959			0,9060
	0,0051	0,6844	0,1133			0,1333		0,8850
	0,0079	0,8748	-0,0499				-0,0133	0,8820
Day 28	28,3029	0,6215	-0,0029					0,8459
	-0,0204	1,4722	-0,6379	0,0261				0,8820
	-0,0253	1,4543	-0,6107		0,0665			0,8870
	0,0192	1,3059	-0,5294			0,0597		0,8350
	0,0095	1,4970	-0,6937				-0,0344	0,8560
Day 90	28,5634	0,7116	-0,0037					0,8654
	-0,0276	1,8278	-0,9080	0,0113				0,9010
	-0,0291	1,8216	-0,8984		0,0249			0,9020
	-0,0098	1,7392	-0,8471			0,0006		0,7052
	-0,0202	1,8481	-0,9367				-0,0326	0,8880

IX. STANDARD DEVIATION

Finally, we created models taking into account the standard deviation by applying a tailor-made version of the Weighted Least Squares method provided by the StatsModels library [11, 12]. The ideal weight is the inverse of the variance of the measurements (more importance is given to the samples with less variability), this is the approach we utilized but it required removing certain points coming from the execution of a single experiment and therefore with standard deviation zero.

We appreciated that the models for later days were more stable and less influenced by this weighted approach than those for earlier days.

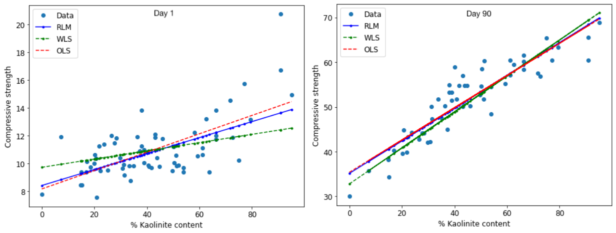


Figure 2: Weighted Least Squares, Ordinary Least Squares, and Recursive Least Squares models for days 1 and 90

After analyzing more in detail the models we observed that another very determining factor for the construction of a method such as this was also to take into account the number of samples that have been used to calculate each mean, since a mean in which only one sample has been used would have a zero standard deviation and yet is a much less reliable sample. Furthermore, for our model to achieve something close to the Best Least Unbiased Estimator (BLUE), many tests or researches were required.

Finally, we analyzed the standard deviation as a function of time to see if the different curve trends obtained in our data are dependent on the days. We obtain a slightly increasing trend until the 27th and it goes down again for the 90th. This is to be expected as with each passing day it is more likely that over time the conditions that each piece of cement analyzed has received are more likely to have been different. However, it is reduced on the last day (day 90) as at this point the cement samples to be tested have been reduced, so the measurements here are a little less reliable.

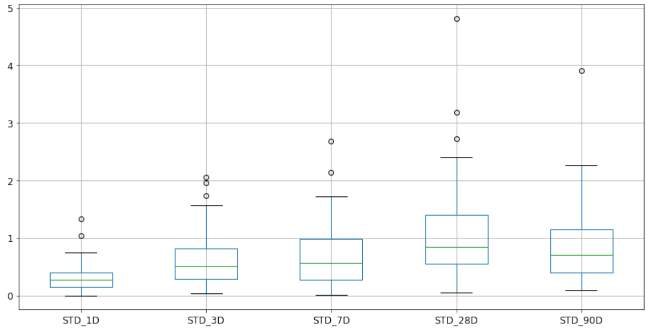


Figure 3: Standard deviations days 1, 3, 7, 28, and 90

X. Conclusions

- The scarcity of data has made us opt for simple models, but which in turn can answer the questions posed to us from the laboratory.
- Kaolinite content has a strong relationship to compressive strength.
- On days 1 and 3 we see a little more random behaviour and a worse performance than in the case of using normal cement, but from day 7 we see a much more stable behaviour. This may be due to the setting time. This is why we consider that it makes more sense to work with the data from the first week, when studying the behavior of clays and making decisions about which clay to use.
- On the 28th and 90th days, we clearly see how for low values of Kaolinite, there is a linear relationship with the compressive strength but for intermediate and high values of Kaolinite the compressive strength tends to stabilise at one value, so the non-linear model is best suited to the data.
- LC3 is a solid alternative to the classical cement in terms of mechanical properties for clays with *Kaolinite Content* greater or equal than 50%.

Acknowledgements

Our deepest thanks to Franco Zunino for allowing us to contribute to this project and help us throughout the process.

We also want to express our gratitude to the Complutense and the Polytechnic Universities of Madrid as well as to the EPFL for all the knowledge and opportunities offered.

References

[1] "Lc3." [Online]. Available: https://lc3.ch/

[2] "Small dataset guidelines." [Online]. Available: https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89

[3] "Linear regression." [Online]. Available: https://medium.com/rants-on-machine-learning/what-to-do-with-small-data-d253254d1a89

[4] "Statsmodel regression." [Online]. Available: https://www.statsmodels.org/stable/regression.html

[5] "Regression with statsmodels ols." [Online]. Available: https://medium.com/python-in-plain-english/ols-linear-regression-basics-with-pythons-scikit-learn

[6] "Understanding statsmodel ols results." [Online]. Available: https://medium.com/@jyotiyadav99111/statistics-how-should-i-interpret-results-of-ols-3bde1ebeec01

[7] "Feature engineering." [Online]. Available: https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features

[8] "Feature selection." [Online]. Available: https://machinelearningmastery.com/an-introduction-to-feature-selection/

[9] "Feature metrics." [Online]. Available: https://machinelearningmastery.com/calculate-feature-importance-with-python/

[10] "Feature selection with scikit learn." [Online]. Available: https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/

[11] "Weighted least squares." [Online]. Available: https://en.wikipedia.org/wiki/Weighted_least_squares

[12] "Weighted least squares with statsmodels." [Online]. Available: https://www.statsmodels.org/stable/examples/notebooks/generated/wls.html