IBRACON Structures and Materials Journal
Revista IBRACON de Estruturas e Materiais
ISSN 1983-4195 ismj.org

REVIEW ARTICLE

Dataset construction and data science analysis of physicochemical characterization of ordinary Portland cement

Construção e análise de banco de dados das propriedades do cimento Portland comum

Jéssica Fall Nogueira Chaves[a]
Francisco Evangelista Junior[a]
João Henrique da Silva Rêgo[a]
Lucas de Paula Vasques[a]

[a]University of Brasilia, Department of Civil and Environmental Engineering, Brasilia, Federal District, Brazil

Received 15 October 2022
Accepted 16 February 2023

Abstract: This paper presents a dataset construction and data science analysis from the literature results of physicochemical characterization of ordinary Portland cement (OPC). The physicochemical variables included the percentage by mass of calcium oxide (CaO), silicon dioxide (SiO_2), aluminum oxide (Al_2O_3), iron oxide (Fe_2O_3), magnesium oxide (MgO), sulfuric oxide (SO_3), sodium oxide (Na_2O), potassium oxide (K_2O), titanium oxide (TiO_2), free lime (CaO_{free}), equivalent alkaline (Na_2O_{eq}), loss on ignition, specific surface, density, water-cement ratio, and compressive strength of cement at 28 days. The searching, collection, and assembly of the dataset aimed to evaluate the information related to those variables through exploratory data analysis, enabling a basic understanding of characterization results of OPCs obtained in publications from different types, sources, years, and countries. The dataset provides a useful source of physicochemical characterization of ordinary cement, and the exploratory data analysis provided an understanding of central, dispersion, and data distribution with statistical metrics of each variable and their pair-wise correlations in the assembled dataset. The constructed dataset and its analysis are a starting point to further data, studies, and artificial intelligence models to provide a broader global view of the production and properties of ordinary Portland cement.

Keywords: Portland cement, oxides, data science, physicochemical properties, compressive strength.

Resumo: Este artigo apresenta a construção de um conjunto de dados e a análise exploratória de dados a partir dos resultados da literatura de caracterização físico-química do cimento Portland comum (CPC). As variáveis físico-químicas incluíram a porcentagem em massa de óxido de cálcio (CaO), dióxido de silício (SiO_2), óxido de alumínio (Al_2O_3), óxido de ferro (Fe_2O_3), óxido de magnésio (MgO), óxido sulfúrico (SO_3), óxido de sódio (Na_2O), óxido de potássio (K_2O), óxido de titânio (TiO_2), cal livre (CaO_{free}), equivalente alcalino (Na_2O_{eq}), perda ao fogo, superfície específica, densidade, relação água-cimento e resistência à compressão do cimento aos 28 dias. A busca, coleta e montagem do conjunto de dados teve como objetivo avaliar as informações relacionadas a essas variáveis por meio de análise exploratória de dados, permitindo uma compreensão básica dos resultados de caracterização de CPCs obtidos em publicações de diferentes tipos, fontes, anos e países. O conjunto de dados fornece uma fonte útil de caracterização físico-química de cimento comum, e a análise exploratória de dados forneceu uma compreensão da distribuição central, de dispersão e de dados com métricas estatísticas de cada variável e suas correlações de pares no conjunto de dados montado. O conjunto de dados construído e sua análise são um ponto de partida para novos dados, estudos e modelos de inteligência artificial para fornecer uma visão global mais ampla da produção e propriedades do cimento Portland comum.

Palavras-chave: cimento Portland comum, óxidos, análise exploratória dos dados, propriedades físico-químicas, resistência à compressão.

How to cite: J. F. N. Chaves, F. Evangelista Junior, J. H. S. Rêgo, and L. P. Vasques, "Dataset construction and data science analysis of physicochemical characterization of ordinary Portland cement," *Rev. IBRACON Estrut. Mater.*, vol. 16, no. 6, e16609, 2023, https://doi.org/10.1590/S1983-41952023000600009

Corresponding author: Jéssica Fall Nogueira Chaves. E-mail: jessicafall@live.com
Financial support: Coordenação de aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Financial Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Universidade de Brasília (FUB) and Decanato de Pós-Graduação (DPG) of the University of Brasília, Brazil.
Conflict of interest: Nothing to declare.
Data Availability: Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data are not available.

1 INTRODUCTION

Portland cement is one of the most used materials worldwide and its compressive strength after 28 days of age is the most used measure concerning engineering and performance properties. This strength is a critical input in the technological control of the material and structural design. At 28 days, it is considered the end of the curing process of cement, and it is expected that the strength specified by the manufacturer will be reached, making the compressive strength a fundamental parameter for comparison, and it is still the most used requirement in the choice of cementitious materials [1], [2]. Therefore, this property is an important criterion for standard compliance and is frequently used in the field of civil construction and scientific research. It is well known that this 28-day compressive strength is influenced by its constituent materials [2], [3].

Ordinary Portland cement (OPC) consists of clinker, which has as raw materials limestone, clay or siliceous materials, and materials containing iron and aluminum oxide, and a small percentage of gypsum to regularize the setting. The cement manufacturing process essentially consists of grinding the raw material, mixing it intimately in certain proportions, and burning (at temperatures of up to around 1.450 °C) in large rotary kilns, where the material is sintered and partially melted. The chemical reaction that takes place between the raw materials of clinker in the rotary kiln generates its four main compounds which are tricalcium silicate or alite (C_3S), dicalcium silicate, or belite (C_2S), tricalcium aluminate (C_3A), and tetracalcium ferroaluminate (C_4AF) [4]–[6]. The proportions of the phases present in the cement influence the physical properties of cementitious materials, such as strength, setting time, among other factors.

It is important to point out that silicates in cement are not pure compounds as they contain secondary oxides in solid solution. Both invariably contain small amounts of magnesium, aluminum, iron, potassium, sodium, and sulfur ions. These oxides exert significant effects on the atomic arrangement, crystal shape and hydraulic properties of silicates. Similar to calcium silicates, Mehta and Monteiro [4] mention that in industrial clinkers both C_3A and C_4AF contain significant amounts of magnesium, sodium, potassium and silica in their crystalline structure [4], [7].

The alkalis, potassium oxide (K_2O) and sodium oxide (Na_2O), because they are soluble and reactive, are among the most common elements in nature and are found in small amounts in all raw materials used in the manufacture of cement, especially in clay compounds. Alkalis are of interest in concrete technology due to their reaction with reactive aggregates, originating the alkali-aggregate reaction that causes disintegration of concrete [8]. However, Neville and Brooks [9] mentions that they influence the speed of development of cement strength.

The Portland Cement manufacturing process has undergone changes to improve the environmental aspect of production. The co-processing technique, for example, is the reusing waste as raw material, or as a source of energy, or both to replace natural mineral resources and fossil fuels such as coal, petroleum, and gas in industrial processes [10]. Although this practice can improve the efficiency of material resources, most waste contains contaminants that can be inserted into some of the secondary oxides in the structure of the 4 main cement compounds and interfere with the products formed. In this way, cleaner production through co-processing therefore requires a good understanding of the impacts of these contaminants on the cement manufacturing process, cement quality and the environment [11], [12].

As much as the present research deals with OPC, commercial cements usually incorporate some type of supplementary cementitious materials (SCM). Some materials such as calcined clay, limestone, silica fume, rice hush ash with controlled burning and metacaulim are studied and increasingly used in the sector as SCM, having different influences on the final product. Limestone, for example, contributes to the process of hydration of cement and during the hydration process there is the formation of carbonamine compounds in the presence of finely ground carbonate material, decreasing the porosity of the cementitious system. In addition, the mechanical strength of cementitious materials is greatly influenced by the presence of SCM, in which the strength gain is slower, being lower in the initial ages and increasing with advancing time [13], [14].

Since the chemical composition of Portland cement and its hydrates have a direct influence on the characteristics of cementitious matrices, its characterization and quantification are of fundamental importance. Quantitative analysis of the concentrations of cement elements is a step widely applied in research that uses it in their experimental programs. Although Portland cement consists essentially of various calcium compounds, the results of routine chemical analyzes are expressed in terms of the elemental oxides present [4].

In research that uses and investigate the OPC, parameters such as its chemical composition, specific surface, density, water-cement ratio, among other properties, are usually investigated together with the results of 28-day compressive strength, since scientific works around the world have already proven the influence between physicochemical properties with the development of mechanical resistance of cementitious materials. The change in the chemical composition of the cement, for example, influences the compressive strength, since the proportions of the different compounds vary significantly from one cement to another. The main oxides, expressed in percentage by mass, investigated in scientific research are calcium oxide (CaO), silicon dioxide (SiO_2), aluminum oxide (Al_2O_3), iron oxide (Fe_2O_3), magnesium oxide (MgO), sulfuric oxide (SO_3), sodium oxide (Na_2O), potassium oxide (K_2O), titanium oxide (TiO_2), free lime (CaO_{free}), and alkaline equivalent (Na_2O_{eq}). The oxide content

of each cement influences the proportion of the main compounds (C_3S, C_2S, C_3A, C_4AF). Because each compound has a different reactivity and forms different products, they influence the mechanical strength in different ways [15]. Several techniques are applied to determine the composition of OPC; however, X-ray fluorescence spectroscopy (XRF) is widely used to characterize the oxides present in Portland cement samples. In addition, X-ray diffraction (XRD) with the Rietveld method and X-ray fluorescence spectroscopy combined with the Bogue Potential calculation are used in studies with ordinary Portland cement to quantify its 4 main compounds (C_3S, C_2S, C_3A, C_4AF).

Regarding the physical characterization, fineness is a parameter used by several researchers that use cementitious matrices, as it is a property that is directly related to the speed of the hydration reaction, having a proven influence on its mechanical behavior. The fineness of the cement is related to the specific surface of the grains and its determination serves mainly to check the uniformity of the material's grinding process. Normally, this property can be measured by using nitrogen adsorption technique (BET), based on a mathematical theory that has the measurement of the specific surface area of a material through the physical adsorption of hydrogen gas molecules on the surface [16], and Blaine air-permeability apparatus, which the specific surface is expressed as area total surface area in square centimeters per gram, or square meters per kilogram, of cement [17]. Since the reaction of Portland cement with water is an effect from the outer surface to the inner surface of the grain, that is, the degree of grinding of the cement will influence the hydration speed and the development of compressive strength [1], [6], [11].

The determination of the compressive strength of Portland cement is standardized worldwide, using cement mortar specimens. The standards of each country establish factors such as dimensions of the specimens, water-cement and sand-cement ratios, type of sand, consistency, among others, to provide uniformity in the process of producing mortars. The American standard C 109/C109M [18] and the British BS EN 196-1 [19], serve as a theoretical basis for the development of standards in different countries. In Brazil, the test method is established by ABNT NBR 7215 [20]. The characterization and mechanical behavior of OPC have been investigated by researchers all over the world with several different goals and results, like Malami et al. [21], Felekoğlu et al. [22], Parande et al. [23], Yao and Sun [24], Dhandapani et al. [25] and Yun et al. [26], however, there is an absolute lack of statistical studies of its oxide components and standard properties. Furthermore, no paper in literature collected and statistically quantified physicochemical characteristics of OPC's composition considering different sources, years, and countries.

This paper aims the collection and analysis of a dataset from the literature on the physicochemical characterization of OPC considering as variables the mass percentage of its oxides: CaO, SiO_2, Al_2O_3, Fe_2O_3, MgO, SO_3, Na_2O, K_2O, TiO_2, CaO_{free}, Na_2O_{eq}, loss on ignition; and the commonly reported physical properties: specific surface, density, water-cement ratio, compressive strength at 28 days. From the collected data, the present work also performs a data science exploratory assessment of each variable, as well as their correlations, through an exploratory data analysis to investigate their statistical moments, distribution characteristics, outlier identification, and statistical correlations among variables that make up the dataset. A bibliometric study of the papers was also carried out, showing the scenario in which, these publications are found, as the most frequently used keyword, year and type of publication, main sources. The main contribution of this paper is the collection and assembly of a novel dataset on which the variables are the commonly reported physical properties and chemical composition of OPCs. Furthermore, the exploratory data analysis provided a basic understanding of central, dispersion, and data distribution with statistical metrics of each variable and their pair-wise correlations in the assembled dataset. This set is crucial to statistical regression, machine learning, and artificial intelligence applications to develop predictive models for the compressive strength based on the physicochemical characteristics of OPC.

2 METHODOLOGY, BIBLIOMETRIC REVIEW, AND DATASET COLLECTION AND ASSEMBLY

The data set was formed from the reading of more than 3.000 scientific productions between March and June 2021 through the Scopus database. To standardize during the entire search, the string "Ordinary Portland Cement" was inserted to limit the results in searches that contained the OPC in the titles, abstracts, and keywords. From the results of the research, the titles, abstracts, and especially the topics of materials and methods were read in all works. The selection of a publication was based on the availability of the results of the OPC characterization tests. In other words, to be selected, the research needed to explicitly provide the mass percentage of, at least, the main four OPC oxides CaO, SiO2, Al2O, FeO2, and the 28-day strength characterization. By having the results of these five variables, especially the compressive strength, the results were collected and added to the dataset collecting, when applicable. After the final screening, the dataset was finally formed from 102 publications.

An initial bibliometric review was carried out to analyze selected scientific productions that used the OPC, using the VOSviewer tool, which provides an interface for viewing and analyzing bibliometric and sociometric networks. The analysis in the software was performed regarding the terms of occurrence of the keywords, applying the full counting method to scan the titles, abstracts, and keywords. This tool employs a visualization method based on the

distance between the nodes of the analyzed network, in which the distance between two nodes approximately indicates the intensity of the relationship between them, thus, the smaller the distance, the greater the intensity of this relationship. Figure 1 presents the bibliometric network extracted from the VOSviewer software, which presents all the terms used in the titles, abstracts, and keywords. The network shows that the most used word is "compressive strength", having 17 occurrences, 66 links, meaning that this term has 66 connections with other words, and a link strength of 76, which indicates the number of publications in which it appears linked to other keywords.

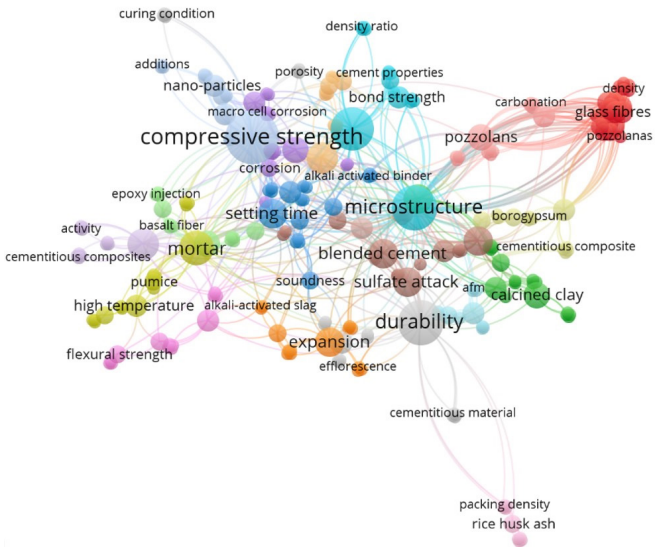


Figure 1. Network of co-occurrence words.

The development of the dataset consisted in the search and collection of the mass percentage values of the main OPC oxides (*CaO*, *SiO₂*, *Al₂O*, *FeO₂*, *MgO*, *SO₃*, *Na₂O*, *K₂O*, *TiO₂*, *CaO$_{free}$*, *Na₂O$_{eq}$*), loss on ignition, specific surface, density, water-cement ratio, and compressive strength at 28 days of the OPCs used in scientific articles published in literature including high-impact journals, books, thesis, and dissertations.

Table 1 presents the physicochemical variables (oxides or properties) considered in this study. Note that the table also specifies how each variable is labeled in all tables and graphs in this manuscript.

Table 1. Physicochemical variables used on the dataset: oxides and test properties.

	Variable	Unity	Label
Oxides	CaO	%	*CaO*
	SiO_2	%	*SiO2*
	Al_2O_3	%	*Al2O3*
	Fe_2O_3	%	*Fe2O3*
	MgO	%	*MgO*
	SO_3	%	*SO3*
	Na_2O	%	*Na2O*
	K_2O	%	*K2O*
	TiO_2	%	*TiO2*
	CaO_{free}	%	*Caofree*
	Na_2O_{eq}	%	*Na2Oeq*
Properties	Loss on Ignition	%	*loss*
	Specific Surface	m^2/kg	*surface*
	Density	g/cm^3	*dens*
	Water-cement ratio	-	*wc*
	28-day compressive strength	MPa	*strength*

Figure 2 presents the evolution of the number of scientific productions published per year among the selected 102 publications. It is possible to notice that from the year 2015 up to 2020 the number of publications increases. The only exceptions are 2019 and 2021, in which this last one was only partially elapsed by the data this paper was written.

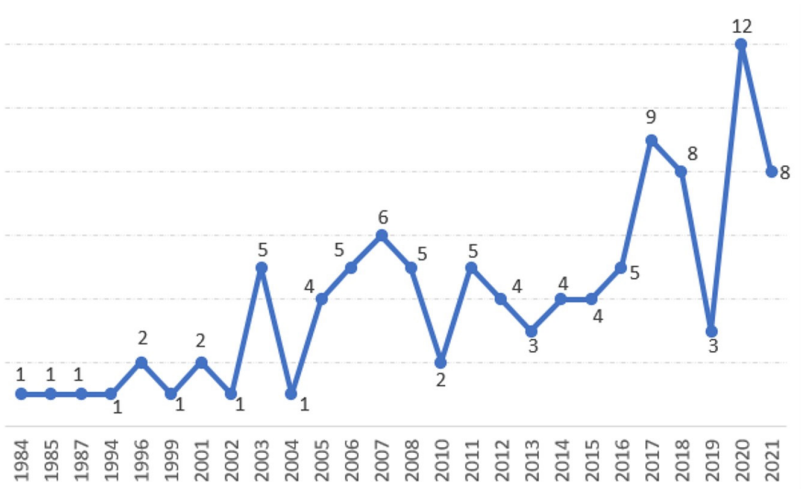


Figure 2. Number of scientific productions per year.

The 102 scientific productions consisted of 92 journal papers (11 international journals), 2 book results (Calcined Clay for Sustainable Concrete), and 8 thesis/dissertations. Figure 3 contains the main sources with their respective numbers of selected publications that characterized the physicochemical properties of OPC. Construction and Building Materials (CBM) stands out with 33 publications, followed by Cement and Concrete Research (CCR) and Cement and Concrete Composites (CCC), with 22 and 17 papers, respectively.

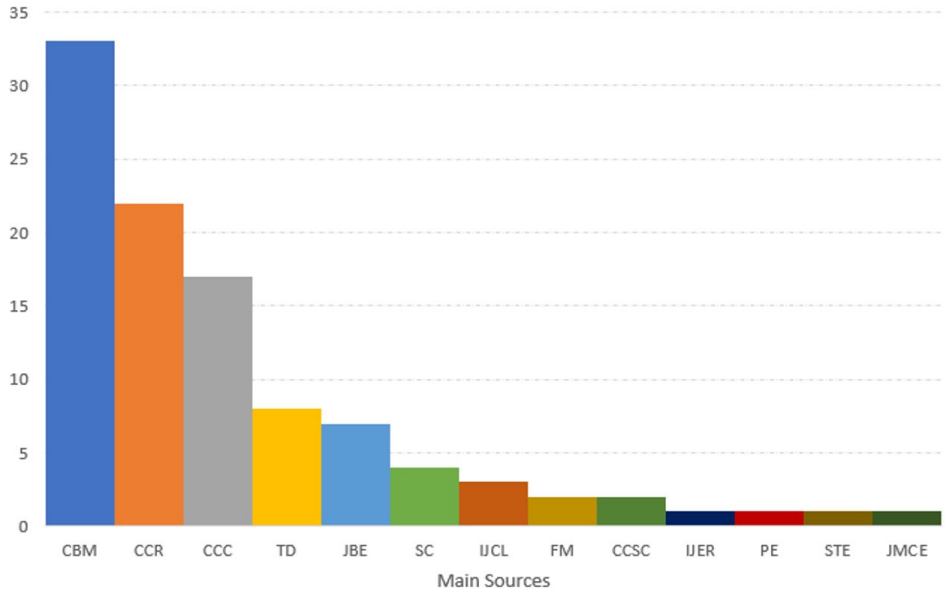


Figure 3. Number of papers per source. Legend: Construction and Building Materials (CBM), Cement and Concrete Research (CCR), Cement and Concrete Composites (CCC), Thesis/Dissertation (TD), Journal of Building Engineering (JBE), Structural Concrete (SC), The International Journal of Cement Composites and Lightweight Concrete (IJCL), Fire and Materials (FM), Calcined clay for Sustainable Concrete (CCSC), International Journal of Energy Research (IJER), Procedia Engineering (PE), Science of the Total Environment (STE) and Journal of Materials in Civil Engineering (JMCE).

The graph plotted in Figure 4 shows the data-filling matrix of the assembled dataset, in which the white blanks represent the lack of data for a given variable, and the blue color represents the presence of data. It is possible to observe that TiO_2 is the parameter that was least provided in the literature, either because it was not investigated in the respective publication or because it was not identified in the chemical characterization test, followed by alkalis (Na_2O and K_2O)

and CaO_{free}. In addition, the sample preparation process for chemical characterization tests, such as XRF, for example, can influence the accuracy of the determination of the percentage of alkalis. The strength was the only parameter that had results shown in all publications. The oxides CaO, SiO_2, Al_2O_3 e Fe_2O_3, which give rise to the four main compounds of Portland cement, are also factors that the researchers sought to investigate and that were described in the characterization of OPC. Likewise, the determination of the specific surface (*surface*) was present in the works and the water-cement ratio (*wc*) was indicated through the standard of the respective country of the publication.

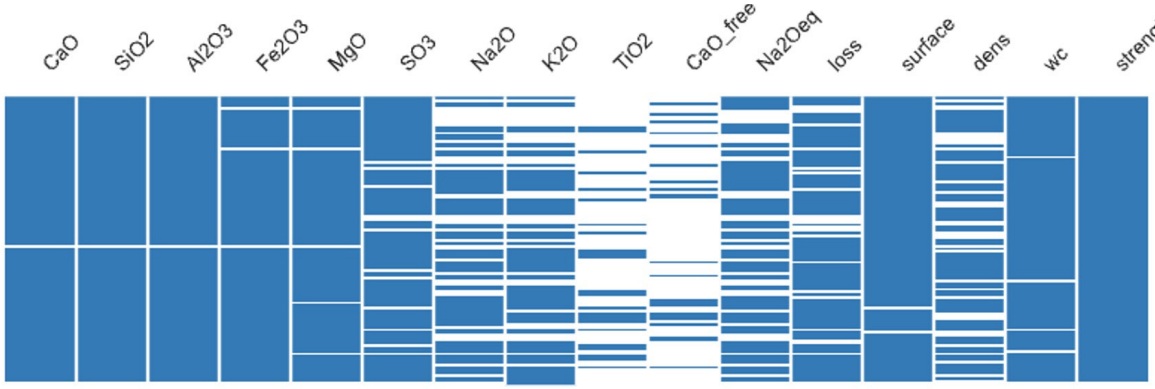


Figure 4. Data-filling matrix.

3 DATA SCIENCE ANALYSIS

From the assembled dataset, an exploratory analysis was performed, which consisted primarily of data preparation, exploratory data analysis of each variable, and the relationships between them. Statistical metrics such as maximum, minimum, mean, and median; metrics of dispersion/variability such as standard deviation, coefficient of variation, range, and outliers detection; and metrics of data distribution such as interquartile range and skewness were also performed. An in-house software Tyche [27] with its data science module Datum was used for those metrics.

The analysis starts with Figure 5 showing a data matrix with pair-wise scatter plots and individual histograms for the 16 variables. The main diagonal shows the individual histograms for each variable, whereas the off-diagonal components, shows the dispersion plot for a combination of two variables. For example, the plot in the first row and the third column is the scatter plot of CaO versus Al_2O_3 oxide variables. Note that the histogram for the water-cement ratio (*wc*) almost represents a categorical variable with only three valid bins: 0.40-0.41, 0.48-0.49, and 0.49-0.50, in which the latter is significatively dominant. This is because almost all standards used the 0.50 water-cement ratio. Fewer exceptions used slightly different values which were counted in the other bins. The Indian standard for determining compressive strength IS 4031 (Part 6) [28], for example, considers that the water-cement ratio is acquired through another standard of the slump measurements IS 4031(Part 4) [29].

Although the water-cement ratio (*wc*) presented this extreme concentration, at the 0.5 ratios, almost as a deterministic variable for this dataset, the authors decided to keep it for completeness of the analysis. The oxides TiO_2 and CaO_{free} also presented a dominant value, but to a lesser degree than *wc*, as showed in the histograms. This is also observed on the scatter plots of those variables (rows 9, 10, and 15) that tend to form horizontal lines on the plots. Figure 5 only allows an initial qualitative assessment of data, therefore, the following paragraphs present an in-depth quantitative analysis of variables. Levels of heading establish the hierarchy of sections by the format or appearance. The section and subsection headings must be preceded by progressive numbering, presented in Arabic numerals, starting at 1.

The histogram of the compressive strength does not suggest any conventional probability distribution. Goodness of fit tests were performed for the main known distributions: normal, log normal, uniform, exponential, extreme value, and they all failed the null-hypothesis showing that there is significant difference between the observed strength values and the expected distributions. The determination of a possible probability distribution will be further investigated in future papers.

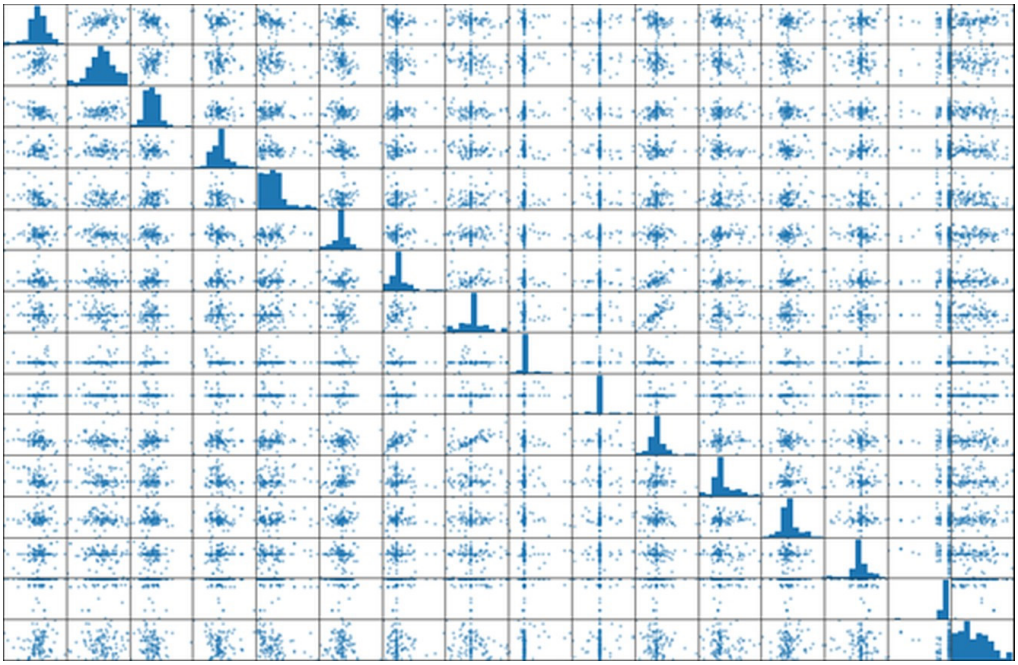


Figure 5. Matrix with histograms and pair-wise scatter plots of the variables.

Table 2 presents the summary of the main statistical parameters for the samples of each variable of the dataset: minimum (min), maximum (max), mean, standard deviation (SD), and coefficient of variation (CV), median, skewness, and interquartile range (IQR).

Table 2. Summary of statistical parameters of each variable.

Variables	Min	Max	Range	Mean	SD	CV	Median	Skewness	IQR
CaO	56.00	68.97	12.97	63.27	2.04	3%	63.44	-0.68	1.90
SiO_2	16.51	23.70	7.19	20.60	1.48	7%	20.65	-0.27	1.88
Al_2O_3	2.45	9.82	7.37	5.06	0.98	19%	5.10	1.02	1.20
Fe_2O3	0.30	7.69	7.39	3.43	1.08	31%	3.34	0.95	1.05
MgO	0.60	6.51	5.91	2.08	1.15	55%	1.93	1.55	1.36
SO_3	0.59	6.24	5.65	2.58	0.80	31%	2.62	0.71	0.77
Na_2O	0.01	0.98	0.97	0.27	0.18	65%	0.22	1.69	0.16
K_2O	0.10	1.32	1.22	0.63	0.30	48%	0.59	0.35	0.43
TiO_2	0.14	0.56	0.42	0.29	0.10	35%	0.24	1.32	0.11
CaO_{free}	0.15	2.41	2.26	1.23	0.66	53%	1.16	0.19	1.03
Na_2O_{eq}	0.08	1.71	1.63	0.65	0.27	42%	0.63	1.00	0.34
Loss	0.00	4.50	4.50	1.79	0.91	51%	1.55	0.70	1.09
Surface	175.00	582.00	407	352.83	59.66	17%	347.00	0.76	51.30
Dens	3.00	3.22	0.22	3.13	0.04	1%	3.13	-0.74	0.05
wc	0.40	0.50	0.10	0.49	0.02	4%	0.50	-3.31	0.01
Strength	38.23	71.00	32.77	50.74	7.90	16%	50.00	0.38	11.83

From the values of the percentage by mass of the oxides from the collected dataset, the mean values of the main components of ordinary Portland cement, C_3S (53,98%), C_2S (17,87%), C_3A (6,92%) e C_4AF (10,15%) were calculated, in percentage, using Bogue's equations [21] considering the addition of gypsum. It is noteworthy that the remainder of the sum of the percentage values of those four compounds was adopted as the content of incorporated calcium sulfate and impurities, determining a mean value of 11.08%.

For the calculation using Bogue, it is necessary to consider that the composition of the four main components of Portland cement are C_3S, C_2S, C_3A and C_4AF with theoretical stoichiometries; all Fe_2O_3 present reacts with Al_2O_3 and CaO to turn into C_4AF; the remaining Al_2O_3 reacts with the CaO to produce C_3A; the remaining CaO reacts with SiO_2

and becomes C_3S and C_2S. The method also considers non-real clinker temperatures close to 2,000 °C, perfect combination of oxides, the existence of balance between C_3S, C_2S and liquid phase [30], [31]. According to Gobbo [30] the calculation restricts the constitution of the cement clinkers to C_3S, C_2S, C_3A e C_4AF, being that it despises the existence of minor elements, such as the TiO_2, MgO, K_2O e Na_2O, among others. It is important to emphasize that some impurities, instead of being present in the cement material, may be incorporated into the structures of main compounds.

3.1 Analysis of sample dispersion, distribution properties, and outliers

The coefficient of variation (CV) conveys the data dispersion (variability of sample data) in terms of the ratio of the standard deviation (SD) and the sample mean values. The CV is a suitable quantity because it expresses the variability of the data excluding the influence of different scales allowing direct comparison among variables of different units or order of magnitude. Figure 6 shows the CV, in percentage, for the 16 variables in descending order. The graph shows that two oxides (*CaO* and SiO_2), *dens*, and *wc* has CV below 8% meaning that their values used around the world to manufacture OPC have very low variability. This was already mentioned for the water-cement ratio since the histogram already showed almost exclusivity of values within the range 0.49-0.50 because of uniformity of compressive strength standards used the 0.5 ratios as discussed before. Another important factor is that *CaO* and SiO_2 are the first and second most dominant components of OPC's mass with means 63.27% and 20.60%, respectively, according to Table 2, which both account for approximately 84% (mean) in OPC's mass. This further showed that those two components mostly related to the manufacturing and extraction process of cement raw materials showed very low variability in their percentage composition in OPCs in the reported literature. However, the CV for the strength was 16%, which is almost double the CV for the most dominant components (in mass). The dashed line in Figure 6 allows a visual comparison of the compressive strength's coefficient of variation with the other variables. Furthermore, the compressive strength samples had a standard deviation of 7.9 Mpa and its mean was 50.74 Mpa, since the values obtained for strength in the research range mainly from 40 to 60 Mpa, having only 9 out of 102 that had resistance below 40 Mpa and only 1 out of 102 above 70 Mpa.

The majority of the other oxides presented high CV, such as SO_3 (31%), Fe_2O (31%), TiO_2 (35%), Na_2O_{eq} (42%), K_2O (48%), CaO_{free} (53%), MgO (55%) e Na_2O (65%). This greater variability in the mass percentage values of these oxides can be explained by the influence of impurities present in the raw materials extracted and used for cement manufacturing, as well as by adjustments made in the chemical composition of the material in each country due to some specific standard. Among the test properties variables, the loss on ignition (*loss*) presented a high CV value (51%) demonstrating the high variability of these test results in the literature.

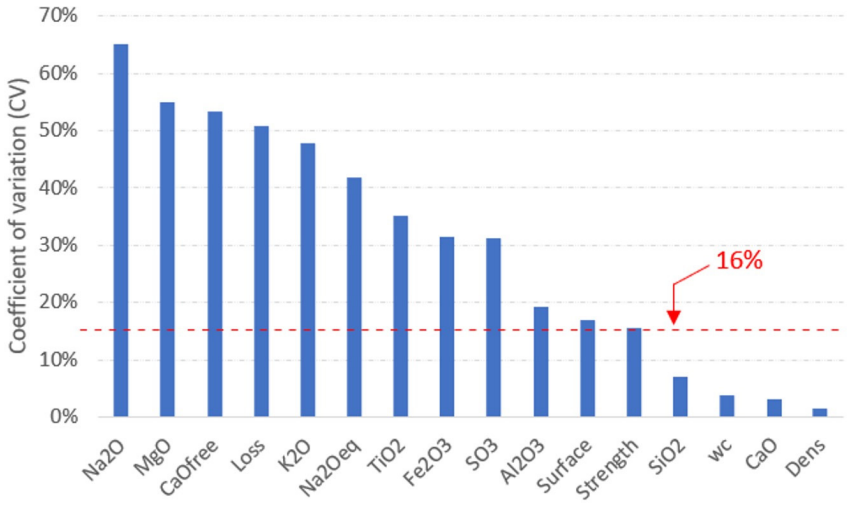
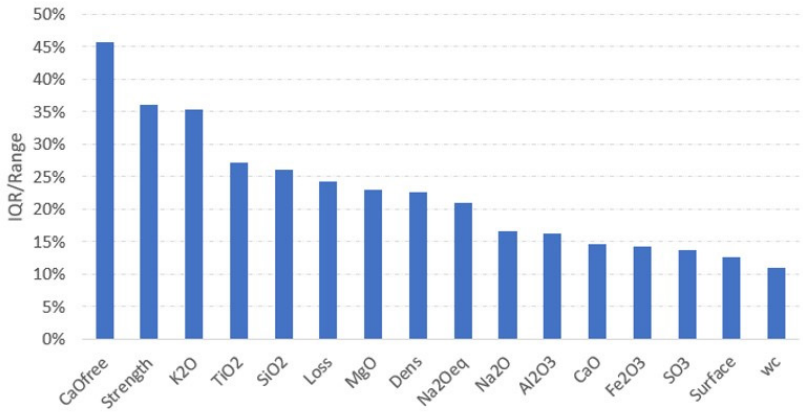


Figure 6. Coefficient of variation (CV) of each variable.

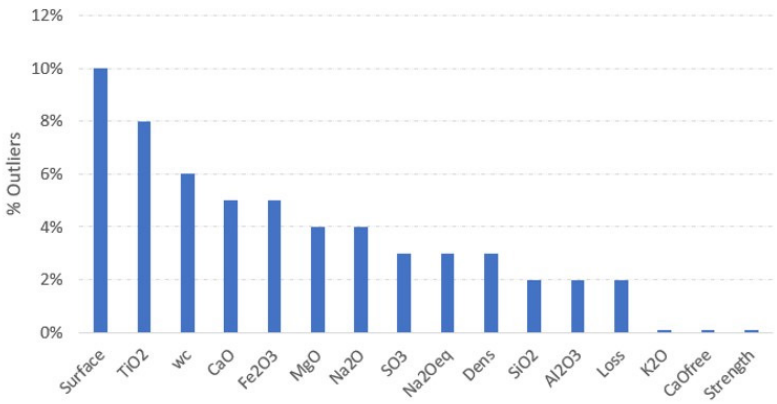
The interquartile range (IQR) measures the data sample distribution in-between the first (Q1) and third (Q3) quartiles (between 25th and 75th percentile). Therefore, IQR shows the range of values around the median (second quartile Q2) corresponding to the 50% central samples. Smaller IQR values imply more sample values toward the left and right tails: lowest and highest 25% values. The IQR and total range (max – min) values for each variable are presented in Table 2.

Figure 7a shows the IQR/Range ratio, in percentage, for each variable. The component CaO_{free} presented the highest ratio showing that 46% of the range amplitude correspond to 50% of the data samples. The K_2O and strength presented slightly more than one-third of range amplitude as IQR. This shows those three variables had quantitatively compacted data distribution around the median as shown in Figure 5. However, the oxides CaO, Fe_2O_3, SO_3; and the property specific surface (*surface*) had less than 15 of their respective total amplitude as IQR which shows that more than 85% of their samples are below the first quartile (25% lowest values) and above the third quartile (25% highest values). The water-cement ratio (*wc*) presented the lowest IQR/Range percentage due to the concentration of values on the right-hand side of the data distribution as shown and explained before in Figure 5.

Based on each IQR, a systematic method of identifying outliers can be used to establish limit values outside Q1 and Q3. The lower limit is 1.5IQR – Q1, while the upper limit is 1.5IQR + Q3, and any sample value outside those limits is considered an outlier. Figure 7b shows the percentage of outliers identified for each variable showing the specific surface, TiO_2, CaO, Fe_2O_3, SO_3 had more than 5% of each respective sample data as outliers. This agrees with the results of Figure 7a, which indicate higher percentages of the range of those variables toward the tails. The only exception is the TiO_2 that although presented a reasonable IQR/Range, had 8% of its data as outliers which demonstrated the use of very discrepant content of TiO_2 in the composition of ordinary cement. One possible explanation is that this oxide is not commonly used as a component of OPC. The high amount of 10% outliers of the surface showed a reasonable percentage of extreme results of those tests to characterize the specific surface presented by the literature for similar cement compositions. This property is related to cement grinding, the greater scatter and outlier percentage identified from collected data shows that this process is carried out in different ways in different countries and can influence the speed of hydration reactions and strength gain in the early ages of the final product.



(a) IQR/Range



(b) Percentage of outliers

Figure 7. Interquartile range (IQR) and outlier quantification: a) IQR/Range and b) the percentage of outliers for each variable.

Regarding the data distribution of each variable, the sample skewness was determined to quantitatively assess the level of asymmetry of the data distribution around its mean. Figure 8 shows the skewness values of each variable and presents the schematic representation of symmetry or asymmetries, in which negative skewness indicates that the tail is on the left side of data distribution, and positive skewness indicates the distribution tail is on the right. Approximately null-value skewness indicates symmetric data distribution as shown in the figure. The oxides K_2O, SiO_2, CaO_{free}, and the compressive strength property had symmetric data distribution due to their very small skewness values (< 0.4). However, the three oxides Na_2O, MgO, and TiO_2 had significant skewness to the right side (skewness > 1.0), quantitatively confirming them to have an asymmetric data distribution. All the other variables showed a moderate to low skewness. The only exception is the water-cement ratio (*wc*) that presented -3.2 skewness due to the data concentration at 0.5 of the uniform standards.

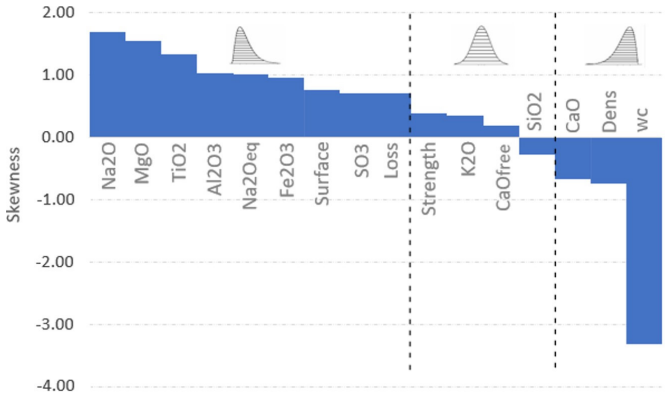


Figure 8. Skewness of data distribution of each variable.

3.2 Correlation between OPC physicochemical properties

The Pearson correlation matrix for the 16 variables is plotted in Figure 9. Due to the symmetry of the matrix, only the lower triangular part is plotted. Each coefficient (ρ) is a measure of linear correlation between two sets of data, and the color intensity means the magnitude of correlation coefficients for pair-wise combinations. Among the oxides, only the correlation between Na_2O_{eq} with Na_2O and K_2O, presented significant positive values of 0.72 and 0.68, respectively. This is somewhat expected since Na_2O_{eq} is derived from the other two oxides. A moderate negative correlation of -0.45 can be observed between CaO and MgO meaning that, when the percentage in the mass of one of these components tends to increase in the cement composition, the other oxide tends to decrease its percentage.

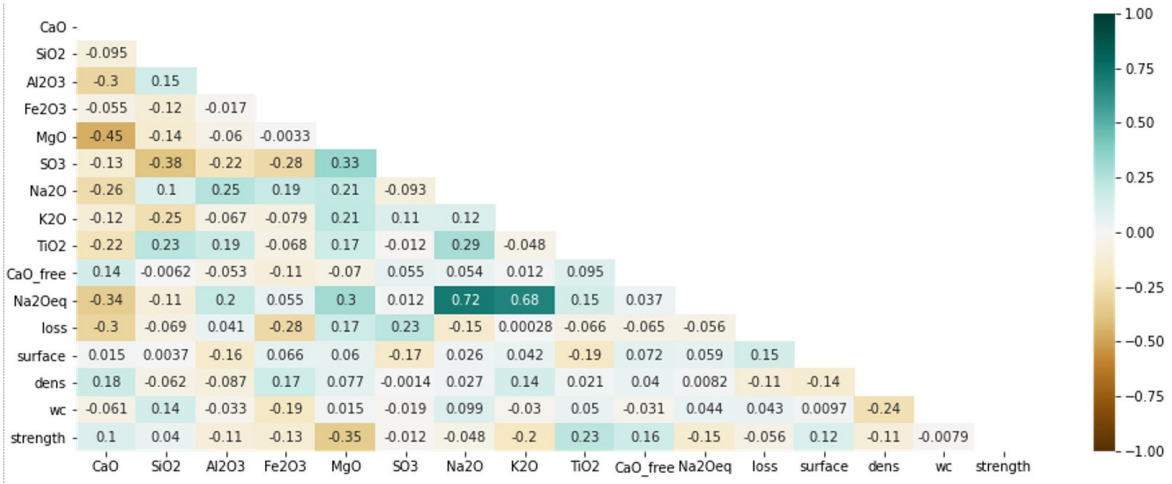


Figure 9. Pearson correlation matrix.

The last row of the Pearson correlation matrix, which corresponds to ρ-values between the compressive strength and all other variables, is plotted in Figure 10. The highest positive correlation (ρ=0.23) was found to be with TiO_2, whereas the highest negative correlation of ρ=-0.35 was with *MgO*. Nevertheless, those magnitudes can be considered moderate to low correlation values. According to Moreno [8], the addition of titanium dioxide to cement aims to adjust the raw material, and *MgO* is derived from the magnesium carbonate present in the original limestone in the form of dolomite, present only in small amounts depending on the specificity of the cement to be produced. Thus, the presence of these two compounds has no evidence of direct influence on compressive strength. It is important to note that the approximately null correlation between the water-cement ratio and compressive strength is due to the very low variability of the collected data set with almost all *wc* values being in 0.50 as already discussed.

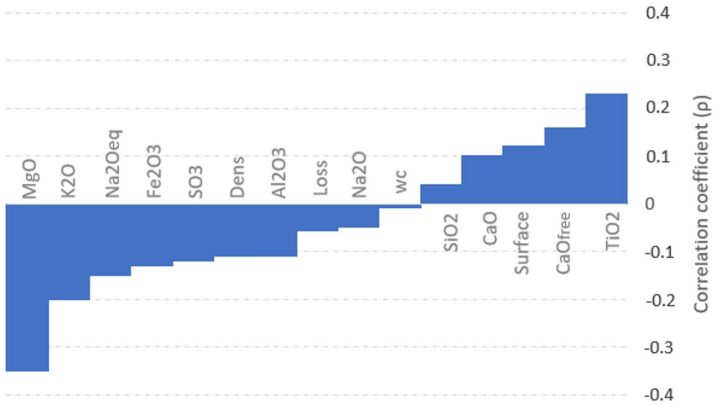


Figure 10. Pearson correlation coefficients between physicochemical variables and OPC's compressive strength.

4 CONCLUSIONS

This paper presented the methodology to construct and analyze a dataset, collected from the literature of different sources, types, and dates with the main thirteen physicochemical properties of Ordinary Portland Cement (OPC) as variables. It was searched the eleven most used oxides components of OPC and five commonly reported properties including the 28-day compressive strength. The dataset was analyzed through an exploratory statistical to quantify the main statistical properties and parameters for each variable and their correlations. The exploratory analysis provided a basic understanding of the data collected and the relationships between the analyzed OPC variables. The constructed dataset is a starting point to further studies with the addition of more data to complement and provide a broader global view of the production and properties of ordinary Portland cement.

The more specific conclusions from the analysis are:

- The main oxides *CaO* and SiO_2 that compose approximately 85% in mass of OPC presented very small data dispersion through their small coefficients of variation (< 7%). Therefore, the composition of OPC produced worldwide has low variability of *CaO* and SiO_2, which had the lowest coefficients of variation and are responsible for 2 of the 4 main components of cement, C_3S, and C_2S, indicating a certain standardization of these compounds. However, the 28-day compressive strength presented a much higher coefficient of variation reaching 16%. Most of the remaining oxides (Na_2O, *MgO*, CaO_{free}, K_2O, Na_2O_{eq}) presented higher dispersion values among the literature in which they had a coefficient of variation greater than 40%. Concerning the C_3A and C_4AF, moderate coefficients of variation were noticed for the oxides Al_2O_3 and Fe_2O_3.
- Compressive strength at 28 days presented a mean value of 50.7 MPa, and the data range for this property is mostly in-between 40 to 60 MPa, but some publications found more scattered values with a minimum value of 38.2 MPa and a maximum of 71.0 MPa. However, the 28-day strength presented symmetric data distribution and 36% of the data were within the IQR. Furthermore, no outlier was detected for the strength data. All these statistical measurements show a compact assemble of the strength for the OPC among the literature, despite the variability and high skewness of the main oxides that compose OPCs.
- Most of the oxides that compose the minority of the OPC in mass had non-symmetric data distribution, especially the Na_2O, *MgO*, TiO_2, and Al_2O_3 presented high skewness to the left (mean greater than the median). Among those

oxides, TiO_2 showed a high value of 8% of the outlier. The specific surface property was the variable that presented the most amount of outliers (10%) showing extreme values for this characterization reported by the literature.

- The ratio between the interquartile range (Q3 – Q1) and the total range (max – min) demonstrated to have a good agreement to the number of outliers detected by each variable, especially variables with higher values for that ratio presented very low or null percentage of outliers on their data samples.
- The correlations showed moderate negative correlation (-0.45) between *MgO* and the main oxide *CaO*, which indicates compositions with higher percentages of MgO had lower percentages of *CaO*. Moreover, the increase in the percentages of *MgO* on the OPC composition, moderately decrease the 28-day compressive strength as indicated by the negative correlation of -0.35 between those variables. The strength did not present any other relevant correlation with other variables.

ACKNOWLEDGEMENTS

This study was partially funded by the Coordenação de aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Financial Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Universidade de Brasília (FUB) and Decanato de Pós-Graduação (DPG) of the University of Brasília, Brazil.

REFERENCES

[1] D. D. S. Mindess and J. F. Young, *Concrete*, 2nd ed. Upper Saddle River: Prentice Hall, 2003.

[2] B. A. Young, A. Hall, L. Pilon, P. Gupta, and G. Sant, "Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods," *Cement Concr. Res.*, vol. 115, pp. 379–388, 2019, http://dx.doi.org/10.1016/j.cemconres.2018.09.006.

[3] T. Oey, S. Jones, J. W. Bullard, and G. Sant, "Machine learning can predict setting behavior and strength evolution of hydrating cement systems," *J. Am. Ceram. Soc.*, vol. 103, no. 1, pp. 480–490, 2020, http://dx.doi.org/10.1111/jace.16706.

[4] P. K. Mehta and P. J. M. Monteiro, *Concrete: Microstructure, Properties, and Materials*, 4th ed. New York: McGraw-Hill Education, 2014. Accessed: Oct. 15, 2022. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9780071797870

[5] W. C. Kosmatka, S. H. Kerkhoff, and B. Panarese, *Design and Control of Concrete Mixtures*, 14th ed. Illinois, USA: Skokie, 2002.

[6] J. W. Bullard et al., "Mechanisms of cement hydration," *Cement Concr. Res.*, vol. 41, no. 12, pp. 1208–1223, 2011, http://dx.doi.org/10.1016/j.cemconres.2010.09.011.

[7] A. M. Neville, *Propriedades do Concreto* [Concrete properties]. Porto Alegre: Bookman, 2016.

[8] F. A. C. Moreno, "Predição da resistência à compressão de um cimento industrial utilizando técnicas de redes neurais artificias" [Prediction of the compressive strength of an industrial cement using artificial neural network techniques], M.S. thesis, Fac. Chem. Eng., State Univ. Campinas, Campinas, 2001.

[9] J. J. Neville and A. M. Brooks, *Tecnologia do Concreto* [Concrete technology], 2ª ed. Porto Alegre: Bookman, 2013.

[10] W. de Q. Lamas, J. C. F. Palau, and J. R. de Camargo, "Waste materials co-processing in cement industry: ecological efficiency of waste reuse," *Renew. Sustain. Energy Rev.*, vol. 19, pp. 200–207, 2013, http://dx.doi.org/10.1016/j.rser.2012.11.015.

[11] A. A. Bogush et al., "Co-processing of raw and washed air pollution control residues from energy-from-waste facilities in the cement kiln," *J. Clean. Prod.*, vol. 254, pp. 119924, 2020, http://dx.doi.org/10.1016/j.jclepro.2019.119924.

[12] M. Achternbosch, K.-R. Bräutigam, N. Hartlieb, C. Kupsch, U. Richers, and P. Stemmermann, "Impact of the use of waste on trace element concentrations in cement and concrete," *Waste Manag. Res.*, vol. 23, no. 4, pp. 328–337, 2005, http://dx.doi.org/10.1177/0734242X05056075.

[13] B. Lothenbach, K. Scrivener, and R. D. Hooton, "Supplementary cementitious materials," *Cement Concr. Res.*, vol. 41, no. 12, pp. 1244–1256, 2011, http://dx.doi.org/10.1016/j.cemconres.2010.12.001.

[14] J. F. N. Chaves, D. L. Nascimento, E. Fonseca, and J. H. S. Rêgo, "Estado da arte sobre o Cimento LC^3," [State of the art about the LC^3 Cement], *Brazilian Congress of Concrete*, 2017.

[15] R. H. Bogue, *The Chemistry of Portland Cement*, 2nd ed. Washington: Reinhold Publishing, 1955.

[16] D. S. Andrade, "Microestrutura de pastas de cimento Portland com nanossílica coloidal e adições minerais altamente reativas [Microstructure of Portland cement pastes with colloidal nanosilica and highly reactive mineral additions]," Ph.D. dissertation, Univ. Brasília, Brasília, 2017.

[17] P. Palacios, M. Hadi, K-K. Mantellato, and S. Bowen, "Laser diffraction and gas adsorption techniques," in *Practical Guide to Microstructural Analysis of Cementitious Materials*, K. L. Scrivener, R. Snellings and B. Lothenbach, Eds., Boca Raton: CRC Press, 2016, pp. 446–472.

[18] American Society for Testing and Materials, *Standard Test Method for Compressive Strength of Hydraulic Cement Mortars*, ASTM C109/C109M-21, 2021.

[19] European Standard, *Methods of Testing Cement - Part 1: Determination of Strength*, EN 196-1, 2016.

[20] Associação Brasileira de Normas Técnicas, *Cimento Portland - Determinação da Resistência à Compressão de Corpos de Prova Cilíndricos*, NBR 7215, 2019.

[21] C. Malami, V. Kaloidas, G. Batis, and N. Kouloumbi, "Carbonation and porosity of mortar specimens with pozzolanic and hydraulic cement admixtures," *Cement Concr. Res.*, vol. 24, no. 8, pp. 1444–1454, 1994, http://dx.doi.org/10.1016/0008-8846(94)90158-9.

[22] B. Felekoğlu, K. Ramyar, K. Tosun, and B. Musal, "Sulfate resistances of different types of Turkish Portland cements by selecting the appropriate test methods," *Constr. Build. Mater.*, vol. 20, no. 9, pp. 819–823, 2006, http://dx.doi.org/10.1016/j.conbuildmat.2005.01.048.

[23] A. K. Parande, B. Ramesh Babu, M. Aswin Karthik, K. K. Deepak Kumaar, and N. Palaniswamy, "Study on strength and corrosion performance for steel embedded in metakaolin blended concrete/mortar," *Constr. Build. Mater.*, vol. 22, no. 3, pp. 127–134, 2008, http://dx.doi.org/10.1016/j.conbuildmat.2006.10.003.

[24] Y. Yao and H. Sun, "Durability and leaching analysis of a cementitious material composed of high volume coal combustion byproducts," *Constr. Build. Mater.*, vol. 36, pp. 97–103, 2012, http://dx.doi.org/10.1016/j.conbuildmat.2012.04.100.

[25] Y. Dhandapani, T. Sakthivel, M. Santhanam, R. Gettu, and R. G. Pillai, "Mechanical properties and durability performance of concretes with Limestone Calcined Clay Cement (LC3)," *Cement Concr. Res.*, vol. 107, pp. 136–151, 2018, http://dx.doi.org/10.1016/j.cemconres.2018.02.005.

[26] C. M. Yun, M. R. Rahman, C. Y. W. Phing, A. W. M. Chie, and M. K. Bin Bakri, "The curing times effect on the strength of ground granulated blast furnace slag (GGBFS) mortar," *Constr. Build. Mater.*, vol. 260, pp. 120622, 2020, http://dx.doi.org/10.1016/j.conbuildmat.2020.120622.

[27] Tyche, *Data Science and Artificial Intelligence Software*. 2020.

[28] Bureau of Indian Standards, *Methods of Physical Tests for Hydraulic Cement Part 6 Determination of Compressive Strength of Hydraulic Cement Other Than Masonry Cement (First Revision)*, IS 4031 (Part 6), 2005, pp. 1–3. Accessed: Oct. 15, 2022. [Online]. Available: https://ia800400.us.archive.org/0/items/gov.in.is.4031.6.1988/is.4031.6.1988.pdf

[29] Bureau of Indian Standards, *Methods of Physical Tests for Hydraulic Cement. Part IV- Determination of Consistency of Standard Cement Paste*, IS 4031 (Part 4), 1988.

[30] L. A. Gobbo, "Aplicação da difração de raios-X e método de Rietveld no estudo de Cimento Portland" [Application of X-ray diffraction and Rietveld Method in the study of Portland cement], M.S. thesis, Inst. Geosci., Univ. São Paulo, São Paulo, 2009.

[31] S. H. Shim, T. H. Lee, S. J. Yang, N. B. M. Noor, and J. H. J. Kim, "Calculation of cement composition using a new model compared to the bogue model," *Materials*, vol. 14, no. 16, pp. 4663, 2021, http://dx.doi.org/10.3390/ma14164663.

Author contributions: J.F.N.C conceptualization, experimental procedure, methodology, analysis, and writing; F.E.J experimental procedure, data curation, supervision and revising; J.H.S.R supervision and revising; L.P.V supervision.

Editors: Fernando Pelisser, Guilherme Aris Parsekian.