Can we really trust traditional compression index correlations? Lessons from a global soil database

Esteban Díaz[1], *Giovanni* Spagnoli[2*], and *Stefano* Collico[3]

[1] Departamento de Ingeniería Civil, Universidad de Alicante, Alicante, Spain
[2] CDM Smith SE, Bochum, Germany
[3] Dipartimento di Ingegneria Civile e Ambientale, Universitá degli Studi di Firenze, Firenze, Italy

Abstract. The compression index (C_c) is a key parameter for predicting consolidation settlement in fine-grained soils. In geotechnical practice, empirical correlations based on simple index properties are widely used as a fast and economical alternative to laboratory oedometer tests. However, most existing formulations were developed under site-specific conditions, raising concerns about their reliability when applied beyond their original scope. In this study, the predictive performance of 20 widely used empirical correlations for C_c estimation is evaluated using a comprehensive global database comprising 1008 soil samples compiled from multiple countries and geological settings worldwide. The dataset spans wide ranges of liquid limit (17.1–199.0%), plasticity index (2.0–82.0%), initial void ratio (0.279–7.114), natural water content (8–244.1%), and C_c (0.013–2.2). Correlations are assessed for the complete dataset and for different compressibility ranges using multiple performance metrics, with Theil's inequality coefficient adopted as the main ranking criterion. Results show that most traditional correlations exhibit large prediction errors and high dispersion when applied globally, particularly for low-compressibility soils. A simple correlation recently proposed by the authors, based solely on natural water content, demonstrates competitive performance, negligible bias, and improved robustness, offering a practical alternative for preliminary C_c estimation in data-scarce conditions.

1 Introduction

Correlations and empirical relationships are widely employed in geotechnical engineering as a rapid and cost-effective way to estimate parameters that are otherwise difficult, time-consuming, or expensive to determine directly. Typically, readily measurable soil properties are related to parameters that require more elaborate laboratory procedures, provided that such correlations are used within their intended range of applicability. Their reliability depends, to varying extents, on soil type, testing procedures, and the degree of deposit uniformity [1]. Consequently, a large number of correlations between soil properties have been proposed, with compilations such as Kulhawy and Mayne [2].

Among the parameters commonly estimated using empirical correlations are properties derived from oedometer tests, which play a key role in geotechnical design due to their direct influence on settlement predictions for structures founded on fine-grained soils. Laboratory oedometer tests provide essential parameters such as the compression index (C_c), which governs the magnitude of consolidation settlement, and the coefficient of consolidation (c_v), which controls the rate at which settlement develops. However, these tests are relatively expensive and time-consuming compared with routine index testing, which can limit their practical application, particularly in small or preliminary projects.

Given these practical constraints, numerous correlations have been proposed to estimate C_c from index properties. One of the earliest and most cited relationships was introduced by Skempton and Jones [3], who related C_c to the liquid limit (LL) for remoulded soils. This contribution stimulated extensive subsequent research, resulting in correlations involving LL, plastic limit (PL), plasticity index (PI), natural water content (w), initial void ratio (e_0), dry unit weight, and other physical descriptors [4-8]. Several investigations have shown that correlations based on w can represent both water bound to clay particles and free pore water, both of which are relevant to compressibility behaviour [6, 9, 10]. Likewise, the e_0 has been recognised as a mechanically meaningful variable, as it captures the in situ state of the soil, including existing stresses and pre-consolidation effects [7, 11, 12]. Other authors have explored multi-parameter formulations, combining variables such as LL with w or with e_0, to enhance predictive capability [9, 13]. Relationships based on PI have also been proposed, acknowledging the influence of clay mineralogy and electrochemical interactions on compressibility [8, 14]. Despite the large number of available empirical correlations, their application to independent datasets often leads to substantial dispersion. When applied outside the conditions under

* Corresponding author: giovanni.spagnoli@cdmsmith.com

which they were derived, deviations of up to 30% have been reported [15], largely because many formulations are based on site-specific or regionally limited datasets that are not representative of different geological contexts. In addition to formulation-related limitations, the C_c itself is affected by inherent experimental uncertainty. Differences between laboratory and field values may result from sampling disturbance, soil fabric anisotropy, specimen size effects, testing procedures, and stress changes induced by excavation. Accordingly, Duncan [16] reported coefficients of variation for C_c ranging from approximately 10% to 37%, indicating that a significant level of scatter is intrinsic to the parameter and is further amplified when empirical correlations are used. Thus, several studies have examined the applicability of existing correlations in specific regions and soil types, which were found to be unsuitable for local conditions [e.g. 17, 18, 19]. These findings indicate that the general applicability of empirical correlations cannot be assumed without prior validation.

Within this framework, the present study performs a global evaluation of widely used empirical correlations for estimating the C_c of fine-grained soils. A large and heterogeneous database compiled from worldwide sources is employed to assess their predictive accuracy and robustness across a broad range of soil conditions. By examining both overall performance and behaviour across different levels of compressibility, this work aims to clarify the limitations of traditional correlations and to provide guidance for their appropriate use in geotechnical practice.

2 Database description

The analysis is based on a comprehensive database of fine-grained soils compiled from published sources and complemented with experimental data obtained by the authors. The final dataset comprises 1008 soil samples, originating from different regions and countries (Nigeria, Ireland, Spain, Iran, Indonesia, Algeria, Bangladesh, Japan and France), each associated with a measured C_c and a set of commonly used index properties (LL, PI, e_0, w). The objective of assembling this database was to ensure a wide coverage of soil types and compressibility levels, enabling the evaluation of empirical correlations under heterogeneous conditions. As summarized in Table 1, LL values range from 17.1% to 199.0%, while PI varies between 2% and 82%, covering soils from low to very high plasticity. The values of e_0 span a broad interval, from 0.279 to 7.114, and w ranges from 8.0% to 244.1%, reflecting markedly different soil structures and moisture conditions.

The C_c values range from 0.013 to 2.200, with a mean value of 0.236, indicating the inclusion of soils from very low to extremely high compressibility. The coefficients of variation (CV) described in Table 1 highlight the pronounced heterogeneity of the dataset, particularly for e_0, w and C_c. This wide dispersion ensures that empirical correlations will be assessed not only within their typical applicability ranges but also under conditions where extrapolation effects may arise.

In addition, the high skewness and kurtosis values reported in Table 1 indicate that most variables exhibit strongly non-normal distributions, characterised by long right tails and the presence of extreme values. This behaviour is particularly evident for e_0, w and c_c, reflecting the high statistical variability of the soils included in the database and the wide range of compressibility conditions represented. In this context, such non-Gaussian distributions support the use of robust performance metrics in the subsequent analysis.

More details about this database can be found in Díaz and Spagnoli [20].

Table 1. Statistical summary of the compiled soil database.

	Mean	Min.	Max.	Skew-ness	Kurto-sis	CV
LL (%)	46.5	17.1	199.0	2.6	16.2	0.34
PI (%)	21.8	2.0	82.0	1.2	4.8	0.45
e_0	0.817	0.279	7.114	5.5	52.3	0.57
w (%)	29.73	8.00	244.1	4.3	32.7	0.57
C_c	0.236	0.013	2.200	4.2	23.1	0.88

Fig. 1 illustrates the relationship between the measured C_c and the main index properties considered in this study: LL, PI, w, and e_0. The joint scatter plots, together with the marginal histograms, provide an overview of the range, dispersion, and distribution of the variables in the database. These plots highlight the strong heterogeneity of the dataset, suggesting approximately linear relationships between C_c and all the considered index properties, which are more pronounced for e_0 and w. It should be noted that LL and PI are influenced by testing procedures and operator interpretation, whereas w is more standardised. The determination of e_0 may also vary depending on the adopted method.

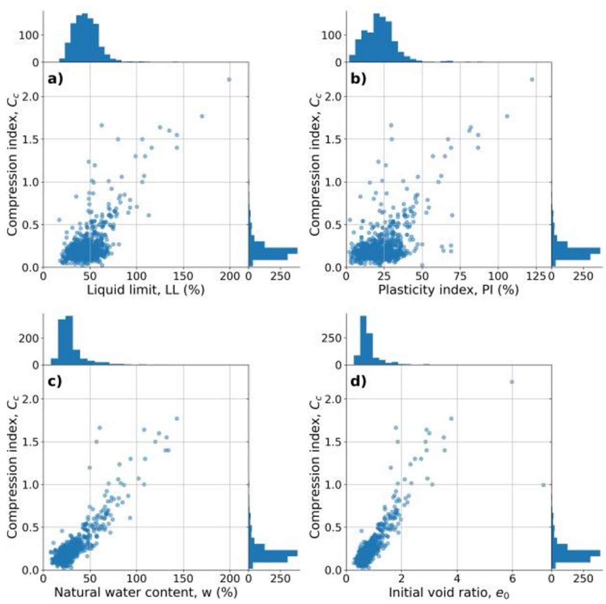

Fig. 1. Scatter plots of measured C_c versus (a) LL, (b) PI, (c) w, and (d) e_0, including marginal histograms.

3 Empirical correlations considered

Numerous empirical correlations have been proposed in the literature to estimate the C_c from basic soil index properties, reflecting decades of experimental evidence obtained under different geological, geographical and depositional conditions. These formulations typically relate C_c to other geotechnical parameters such as LL, w, PI or e_0, and have been widely used in geotechnical practice due to their simplicity and minimal data requirements. In this study, the empirical correlations listed in Table 2 were selected for evaluation. The selection was based on their broad applicability, i.e., correlations developed for general use rather than for specific sites, as well as on their frequent use in the literature and their reliance on commonly available index properties.

4 Evaluation methodology

The empirical correlations are evaluated by comparing predicted and measured values of the C_c for the complete dataset and for several predefined ranges. This range-based analysis allows the performance of each correlation to be examined under different levels of soil compressibility.

The evaluation is based on error metrics with a clear physical interpretation. The Mean Absolute Percentage Error (MAPE) is used to quantify the average relative prediction error, while Theil's Inequality Coefficient (TIC) is adopted as the main ranking criterion. TIC is a scale-free indicator that shows how accurate the predictions are. It separates the total prediction error into three parts: bias, variance, and covariance. Lower TIC values (closer to zero) indicate better model accuracy. Systematic prediction tendencies are assessed using the Mean Error (ME), with positive values indicating overestimation of C_c and negative values indicating underestimation. The dispersion of the estimates is evaluated using the coefficient of variation of the prediction ratio (CV(K)), where K is defined as the ratio between predicted and measured C_c values. The Mean Absolute Error (MAE) is reported as a conventional goodness-of-fit indicator for completeness, although it is not used as the main criterion for performance evaluation.

5 Results and discussion

The evaluation of empirical correlations over the complete dataset reveals pronounced variability in predictive accuracy, characterized by large relative errors and systematic bias effects.

The lowest TIC values are obtained by water-content-based formulations: Nagaraj and Srinivasa Murthy [24] (TIC = 0.152, MAPE = 37.1%) and Azzouz, et al. [6] Eq. 12 (TIC = 0.157, MAPE = 34.7%). Despite ranking as best performers, these correlations still exhibit substantial dispersion (CV(K) ≈ 0.55) and show slight positive bias (ME ≈ 0.01–0.02), indicating a tendency toward overestimation across the dataset.

Table 2. Correlations evaluated in this work.

Eq.	Equation	Soil type / conditions	Reference
(1)	$C_c = 0.009(LL - 10)$	Clayey soils normally consolidated	Terzaghi, et al. [5]
(2)	$C_c = 0.0083(LL - 9)$	Various clays	Schofield and Wroth [21]
(3)	$C_c = 0.006(LL - 9)$	All clays with LL< 100%	Azzouz, et al. [6]
(4)	$C_c = 0.0092(LL - 13)$	Various clays	Mayne [22]
(5)	$C_c = 0.008(LL - 12)$	All clays	Sridharan and Nagaraj [8]
(6)	$C_c = PI/74$	All clays	Wroth and Wood [14]
(7)	$C_c = 0.0054(2.6w - 35)$	Undisturbed clays	Nishida [11]
(8)	$C_c = 0.0102(w - 9.15)$	Inorganic cohesive soils	Hough [12]
(9)	$C_c = 0.01(w - 7.549)$	All clays	Rendon-Herrero [10]
(10)	$C_c = 0.01w$	Chicago and Alberta normally consolidated clays	Koppula [9]
(11)	$C_c = 0.0115w$	Soft clays	Moran, et al. [23]
(12)	$C_c = 0.01(w - 5)$	Clayey soils in USA and Greece	Azzouz, et al. [6]
(13)	$C_c = 0.156e_o + 0.0107$	Cohesive and inorganic soils	Hough [12]
(14)	$C_c = 0.4049(e_o - 0.3216)$	Inorganic cohesive soils	Hough [12]
(15)	$C_c = 0.3(e_o - 0.27)$	All soil types	Rendon-Herrero [10]
(16)	$C_c = -0.156 + 0.41e_o + 0.00058w$	Clayey soils in USA and Greece	Azzouz, et al. [6]
(17)	$C_c = 0.37(e_o + 0.003LL - 0.34)$	Clayey soils in USA and Greece	Azzouz, et al. [6]
(18)	$C_c = 0.009w + 0.005LL$	All clays	Koppula [9]
(19)	$C_c = 0.009w + 0.002LL - 0.10$	Normally consolidated clays	Nagaraj and Srinivasa Murthy [24]
(20)	$C_c = -0.156 + 0.411e_o + 0.00058LL$	All clays	Al-Khafaji and Andersland [13]

A second-tier group, i.e. Rendon-Herrero [10], Azzouz, et al. [6] Eq. 17, and Al-Khafaji and Andersland [13], demonstrates comparable or marginally better stability (MAPE = 26–31%, TIC = 0.16–0.18). These formulations systematically underestimate C_c (negative ME: −0.01 to −0.03) but exhibit lower bias magnitudes than the top-tier group.

Traditional LL-based correlations Terzaghi, et al. [5], Schofield and Wroth [21] and Wroth and Wood [14] show markedly inferior performance (MAPE ≥ 80%, TIC ≈ 0.27–0.29, large CV(K)) with consistent overestimation (ME ≈ +0.07 to +0.09), reflecting strong amplification of experimental scatter.

The most unfavourable behaviour is observed for the formulation proposed by Koppula [9] based on w and LL (Eq. 18), for which prediction errors exceed 160% in relative terms. This correlation presents extremely high dispersion and a strong positive bias (ME ≈ +0.26), indicating severe overestimation of C_c values across the dataset. Overall, the results demonstrate that, when applied to a large and heterogeneous database, empirical correlations yield prediction errors that frequently reach several tens of percent and, in extreme cases, exceed 100%, with both overestimation and underestimation depending on the formulation considered. A summary of the performance metrics for the complete dataset is presented in Table 3, while the corresponding scatter plots comparing measured and predicted values of the best four correlations are shown in Fig. 2. These four correspond to the best-performing correlations for the present dataset.

Table 3. Performance metrics of empirical correlations considering the complete dataset. Empirical correlations are ranked based on TIC.

Eq.	MAE	MAPE (%)	TIC	CV(K)	ME
(19)	0.064	37.124	0.152	0.548	0.023
(12)	0.062	34.699	0.157	0.551	0.010
(9)	0.060	30.629	0.164	0.557	-0.016
(17)	0.053	26.442	0.164	0.524	-0.009
(7)	0.072	37.644	0.165	0.614	-0.007
(8)	0.062	30.698	0.169	0.564	-0.027
(10)	0.087	53.499	0.172	0.553	0.060
(20)	0.059	28.533	0.176	0.541	-0.030
(14)	0.061	29.077	0.185	0.542	-0.036
(16)	0.065	31.490	0.185	0.560	-0.040
(11)	0.119	73.157	0.199	0.553	0.105
(15)	0.082	34.027	0.251	0.530	-0.073
(5)	0.115	65.762	0.263	0.711	0.037
(4)	0.130	78.329	0.263	0.717	0.070
(2)	0.131	79.718	0.266	0.696	0.073
(1)	0.141	87.194	0.270	0.701	0.090
(6)	0.143	85.634	0.287	0.832	0.071
(3)	0.101	48.908	0.294	0.696	-0.014
(13)	0.105	37.526	0.326	0.541	-0.100
(18)	0.268	162.740	0.382	0.580	0.263

To identify systematic patterns in correlation behavior, the dataset was subdivided by C_c ranges, revealing that predictive performance is highly dependent on soil compressibility level. This range-based analysis exposes critical limitations invisible when examining complete-dataset metrics alone. For low C_c values ($C_c < 0.10$), all evaluated empirical correlations exhibit extremely poor predictive performance (Table 4), characterised by very large relative errors, high dispersion, and systematic bias. Even the best-performing formulation, proposed by Hough [12] based on e_0 (Eq. 13), shows a TIC of approximately 0.23 and a MAPE close to 67%, together with very high scatter (CV ≈ 0.72) and a positive ME, indicating systematic overestimation. Most correlations perform very poorly, with MAPE frequently exceeding 100% and reaching several hundred percent in extreme cases, accompanied by high TIC values, large dispersion, and strong positive bias. These results indicate that, in the low-compressibility range, prediction errors often exceed the magnitude of the measured C_c itself, leading to a complete loss of predictive reliability.

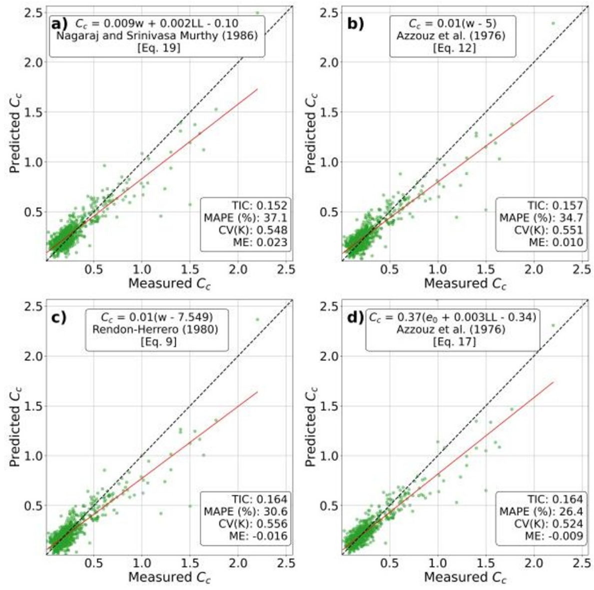

Fig. 2. Comparison between measured and predicted C_c values for the four empirical correlations with the lowest TIC values over the complete dataset. Each panel includes the corresponding formulation, linear regression fit (red line), the 1:1 reference line (dashed), and the associated error metrics.

Table 4. Performance metrics of empirical correlations for C_c < 0.10. Empirical correlations are ranked based on TIC.

Eq.	MAE	MAPE (%)	TIC	CV(K)	ME
(13)	0.033	66.922	0.230	0.721	0.024
(15)	0.042	75.913	0.297	0.886	0.013
(14)	0.054	95.894	0.354	0.955	0.023
(20)	0.055	97.642	0.358	0.975	0.024
(16)	0.055	93.588	0.368	1.030	0.012
(8)	0.062	114.281	0.369	0.880	0.040
(17)	0.064	119.978	0.372	0.868	0.052
(9)	0.070	128.250	0.383	0.841	0.054
(7)	0.070	120.657	0.410	1.033	0.026
(12)	0.086	158.327	0.422	0.796	0.079
(19)	0.093	170.189	0.438	0.807	0.090
(3)	0.117	200.032	0.492	0.794	0.115
(10)	0.129	229.061	0.501	0.744	0.129
(11)	0.160	278.420	0.548	0.744	0.160
(5)	0.156	261.967	0.564	0.822	0.154
(2)	0.188	312.603	0.599	0.794	0.188
(4)	0.181	301.510	0.600	0.833	0.179
(1)	0.201	333.364	0.616	0.802	0.201
(6)	0.197	332.528	0.625	0.977	0.194
(18)	0.311	515.249	0.686	0.718	0.311

For the intermediate-low compressibility range ($0.10 \le C_c < 0.20$), predictive performance improves compared with the $C_c < 0.10$ interval, although errors and dispersion remain high (Table 5). The best results are obtained by the formulation of Azzouz, et al. [6] based on e_0 and LL (Eq. 17), which shows the lowest MAPE (≈ 22%), a relatively low TIC (≈ 0.13), and moderate dispersion (CV ≈ 0.27), together with a small positive ME indicating slight overestimation. Similar behaviour is observed for water-content-based correlations by Rendon-Herrero [10] and Hough [12], with MAPE values around 26–27%, TIC close to 0.15–0.16, and limited bias. The formulation by Nagaraj and Srinivasa Murthy [24] exhibits higher relative error (MAPE ≈ 35%) and a clearer tendency to overestimate. The worst results correspond to Koppula [9], for which relative errors become very large (MAPE up to and above 180%), dispersion increases substantially, and strong positive bias is observed, indicating unstable and unreliable predictions in this range.

Table 5. Performance metrics of empirical correlations for $0.10 \leq C_c < 0.20$. Empirical correlations are ranked based on TIC.

Eq.	MAE	MAPE (%)	TIC	CV(K)	ME
(17)	0.033	22.075	0.131	0.271	0.008
(9)	0.039	26.255	0.151	0.320	0.003
(20)	0.039	25.730	0.162	0.347	-0.018
(12)	0.046	31.423	0.162	0.284	0.029
(8)	0.040	26.724	0.162	0.351	-0.010
(14)	0.040	26.169	0.167	0.352	-0.022
(19)	0.051	35.002	0.173	0.250	0.042
(13)	0.044	27.429	0.183	0.199	-0.043
(16)	0.045	30.054	0.193	0.399	-0.031
(15)	0.046	29.580	0.194	0.307	-0.041
(7)	0.054	35.585	0.215	0.481	-0.016
(3)	0.068	46.373	0.227	0.344	0.052
(10)	0.080	55.066	0.233	0.242	0.079
(11)	0.114	77.387	0.292	0.242	0.114
(5)	0.106	71.839	0.308	0.370	0.097
(2)	0.135	91.159	0.349	0.344	0.131
(4)	0.132	89.044	0.353	0.381	0.125
(1)	0.149	100.890	0.371	0.352	0.146
(6)	0.142	95.327	0.374	0.427	0.129
(18)	0.272	182.730	0.475	0.202	0.272

For the medium compressibility range ($0.20 \leq C_c < 0.50$), predictive performance improves relative to lower C_c intervals, although dispersion and systematic bias remain relevant (Table 6). The best results are obtained with the formulation of Azzouz, et al. [6] based on e_0 and LL (Eq. 17), which provides the lowest MAPE (≈ 18.6%), the lowest TIC (≈ 0.12), and relatively low dispersion (CV ≈ 0.24), together with a small negative ME indicating slight underestimation. A second group of correlations, including Nagaraj and Srinivasa Murthy [24], Azzouz, et al. [6] based on w (Eq. 12), and Al-Khafaji and Andersland [13], shows slightly higher errors (MAPE ≈ 21–23%) and increased dispersion (CV ≈ 0.26–0.28), with bias varying in sign depending on the formulation. The poorest results correspond to the formulations by Koppula [9] based on w and LL (Eq. 18), for which relative errors exceed 95% accompanied by high TIC value and strong systematic bias, indicating unstable and unreliable predictions within this range.

For high compressibility soils ($C_c \geq 0.50$), the predictive performance of the empirical correlations remains heterogeneous, with moderate absolute errors but significant relative errors, dispersion, and systematic bias (Table 7). The best results are obtained for the formulations by Moran, et al. [23], Nishida [11], Koppula [9] based on w (Eq. 10), and Nagaraj and Srinivasa Murthy [24], which show the lowest MAPE values (approximately 16–18%), low TIC values (≈ 0.13–0.15), and relatively limited dispersion (CV ≈ 0.21–0.23). Bias is generally small, although some formulations exhibit slight overestimation (ME ≈ +0.01) while others show moderate underestimation (ME between −0.10 and −0.12).

Table 6. Performance metrics of empirical correlations for $0.20 \leq C_c < 0.50$. Empirical correlations are ranked based on TIC.

Eq.	MAE	MAPE (%)	TIC	CV(K)	ME
(17)	0.052	18.571	0.119	0.242	-0.019
(19)	0.059	21.797	0.128	0.260	0.013
(12)	0.060	22.190	0.134	0.285	0.003
(20)	0.062	22.718	0.143	0.284	-0.037
(9)	0.063	23.081	0.145	0.314	-0.023
(14)	0.065	23.501	0.148	0.284	-0.044
(10)	0.075	28.627	0.149	0.244	0.053
(8)	0.067	24.822	0.157	0.336	-0.034
(16)	0.069	25.172	0.157	0.309	-0.046
(5)	0.074	27.533	0.163	0.339	-0.002
(2)	0.079	30.233	0.167	0.318	0.034
(4)	0.083	31.820	0.177	0.347	0.031
(1)	0.088	34.017	0.181	0.324	0.051
(7)	0.084	30.998	0.184	0.412	-0.003
(3)	0.077	26.554	0.185	0.318	-0.053
(11)	0.113	42.690	0.198	0.244	0.102
(6)	0.099	37.623	0.209	0.418	0.031
(15)	0.092	33.008	0.214	0.259	-0.089
(13)	0.126	44.423	0.312	0.185	-0.126
(18)	0.252	95.818	0.323	0.194	0.252

The poorest results correspond to Hough [12] based on e_0 (Eq. 13), for which MAPE values exceed 60%, together with high TIC values (up to ≈ 0.46), large dispersion, and strong negative bias (ME ≈ -0.52). These results indicate unstable predictions and severe underestimation of the C_c for highly compressible soils.

Table 7. Performance metrics of empirical correlations for $C_c \geq 0.50$. Empirical correlations are ranked based on TIC.

Eq.	MAE	MAPE (%)	TIC	CV(K)	ME
(11)	0.148	16.958	0.125	0.211	0.010
(7)	0.160	18.069	0.138	0.234	0.009
(10)	0.151	15.640	0.142	0.211	-0.101
(19)	0.164	17.718	0.149	0.221	-0.123
(18)	0.268	35.196	0.159	0.220	0.205
(12)	0.177	18.798	0.161	0.216	-0.151
(9)	0.196	21.192	0.172	0.220	-0.176
(8)	0.199	21.724	0.173	0.223	-0.179
(17)	0.200	21.580	0.185	0.308	-0.139
(20)	0.200	21.734	0.190	0.332	-0.127
(16)	0.203	22.076	0.192	0.336	-0.130
(14)	0.218	23.793	0.200	0.339	-0.157
(1)	0.265	28.888	0.228	0.326	-0.247
(4)	0.278	30.739	0.236	0.337	-0.262
(2)	0.296	32.356	0.252	0.323	-0.285
(6)	0.304	33.964	0.259	0.390	-0.288
(15)	0.348	39.503	0.277	0.328	-0.319
(5)	0.335	37.324	0.283	0.333	-0.329
(3)	0.440	49.903	0.380	0.323	-0.439
(13)	0.521	60.118	0.465	0.282	-0.517

These tables provide a quantitative indication of the magnitude of the prediction errors associated with each empirical correlation, which may be useful when selecting a correlation for preliminary estimation in new or data-scarce areas. They enable practitioners to choose correlations based on prioritised metrics (TIC for overall accuracy, ME for bias, and CV(K) for dispersion), depending on specific project requirements. Since all formulations are evaluated using a large and heterogeneous database, the reported error metrics offer a testing framework that is more representative than the limited datasets originally used for the calibration of individual correlations, allowing a more informed and comparative assessment of their expected performance.

The results clearly highlight the high dispersion and the large errors that arise when empirical correlations are applied to datasets outside the conditions under which they were originally calibrated. This limitation should be carefully considered when selecting a correlation for practical applications. To address this issue, and in the absence of more advanced approaches such as machine learning, Díaz and Spagnoli [20] proposed a simple empirical correlation specifically calibrated using the present database, which provides an alternative for estimating C_c:

$$C_c = 0.011w - 0.0905 \quad (21)$$

The proposed correlation based on w exhibits an overall performance comparable to the best-performing formulations identified for the complete dataset analysed in this study. Table 8 summarises the performance metrics of Equation 21 across the different C_c intervals considered.

Table 8. Performance metrics of the Díaz and Spagnoli [20] correlation (Eq. 21) for the complete dataset and different c_c ranges.

	MAE	MAPE (%)	TIC	CV(K)	ME
All data	0.061	32.998	0.153	0.559	-0.001
$C_c < 0.10$	0.075	138.762	0.403	0.856	0.059
$0.10 \leq C_c < 0.20$	0.043	29.263	0.162	0.332	0.011
$0.20 \leq C_c \leq 0.50$	0.065	24.329	0.147	0.323	-0.005
$C_c > 0.50$	0.162	17.362	0.148	0.221	-0.117

Overall, the proposed equation shows a competitive global performance (MAE = 0.061, TIC = 0.153), comparable to the correlations that yield the best results for the complete dataset, such as the formulations by Azzouz, et al. [6] based on e_0 and LL (Eq. 17) and by Nagaraj and Srinivasa Murthy [24]. A key advantage of the proposed equation is its practically null bias (ME ≈ 0), together with the exclusive use of the w, which is a more readily available parameter than combined index properties. When analysed by compressibility ranges, the equation shows particularly good performance for highly compressible soils (C_c >0.50) with performance metrics comparable to those obtained by the best-performing correlations in this range (Moran, et al. [23], Nishida [11]), and maintains stable behaviour for intermediate compressibility levels. As observed for all evaluated correlations, predictive performance deteriorates markedly for low compressible soils (C_c <0.10) where relative errors remain high.

Fig. 3 illustrates the relationship between measured and predicted C_c values obtained using the correlation proposed by Díaz and Spagnoli [20], showing the associated scatter and the deviation from the ideal 1:1 agreement line

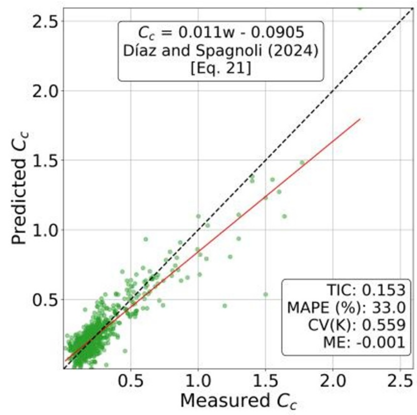

Fig. 3. Comparison between measured and predicted C_c values using the empirical correlation proposed by Díaz and Spagnoli [20] (Eq. 21). The dashed line represents the 1:1 reference line whereas the red line represents the linear regression fitted to the data.

In summary, this equation emerges as a robust and practical alternative for general application, offering a favourable balance between accuracy, bias control, and simplicity.

6 Conclusions

This study evaluated the predictive performance of 20 widely used empirical correlations for estimating the C_c of fine-grained soils using a comprehensive global database of 1008 samples. The correlations were assessed across the complete dataset and within distinct compressibility ranges to examine their accuracy, bias, and stability under heterogeneous soil conditions. The main findings are summarised as follows:

- Most empirical correlations exhibit significant prediction errors when applied to data outside their original calibration conditions. Even the best-performing formulations show MAPE values between 30 and 37% for the complete dataset, while several correlations exceed 100% error. This clearly indicates that site-specific correlations cannot be reliably extrapolated to global applications without a substantial loss of accuracy.
- For the complete dataset, the best results are obtained by Nagaraj and Srinivasa Murthy [24] and Azzouz, et al. [6] based on w (Eq. 12), with MAPE ≈ 30–37% and TIC ≈ 0.15–0.16, although dispersion remains high (CV(K) ≈ 0.55).
- For the low-compressibility range (C_c < 0.10), no correlation provides acceptable predictions; even the least unfavourable case Hough [12] (Eq. 13) based on e_0 shows MAPE ≈ 67% and very high dispersion, indicating loss of reliability.
- For the range $0.10 \leq C_c < 0.20$, Azzouz, et al. [6] based on e_0 and LL (Eq. 17) performs best, with MAPE ≈ 22%, TIC ≈ 0.13, and moderate dispersion, while most other correlations show rapidly increasing errors.

- In the intermediate compressibility range ($0.20 \le C_c < 0.50$), Azzouz, et al. [6] based on e_0 and LL (Eq. 17) remains the best, achieving MAPE ≈ 19% and the lowest TIC, with only slight underestimation.
- For highly compressible soils ($C_c \ge 0.50$), best performance is obtained by Moran, et al. [23], Nishida [11], Koppula [9] based on w (Eq. 10), and Nagaraj and Srinivasa Murthy [24], with MAPE ≈ 16–19% and relatively low dispersion.
- The empirical correlation proposed by Díaz and Spagnoli [20] (Eq. 21) shows an overall performance comparable to the best-performing formulations, while maintaining a practically null bias and relying exclusively on w. It is particularly competitive for the complete dataset and for highly compressible soils.
- This study delivers a set of summary tables that provide guidance on the applicability of existing empirical correlations, enabling informed selection based on soil compressibility and observed prediction performance.

These findings emphasize the critical importance of validating empirical correlations against representative local datasets before their application in geotechnical design.

References

1. M. Uzielli, S. Lacasse, F. Nadim, K. Phoon, Soil variability analysis for geotechnical practice. In Characterisation and engineering properties of natural soils, Vol. 3, 1653–1752 (2007).
2. F.H. Kulhawy, P.W. Mayne, Manual on estimating soil properties for foundation design (Electric Power Research Institute, Palo Alto, CA; Cornell University, Ithaca, 1990).
3. A.W. Skempton, O. Jones, Notes on the compressibility of clays. Q. J. Geol. Soc. **100**, 119–135 (1944).
4. V. Cozzolino, Statistical forecasting of compression index. In Proc. 5th Int. Conf. on Soil Mechanics and Foundation Engineering, Paris (1961), pp. 51–53.
5. K. Terzaghi, R.B. Peck, G. Mesri, Soil mechanics in engineering practice, 2nd ed. (John Wiley and Sons, New York, 1967).
6. A.S. Azzouz, R.J. Krizek, R.B. Corotis, Regression analysis of soil compressibility. Soils Found. **16**, 19–29 (1976). https://doi.org/10.3208/sandf1972.16.2_19
7. J.E. Bowles, Physical and geotechnical properties of soils (McGraw-Hill, New York, 1979).
8. A. Sridharan, H. Nagaraj, Compressibility behaviour of remoulded, fine-grained soils and correlation with index properties. Can. Geotech. J. **37**, 712–722 (2000).
9. S. Koppula, Statistical estimation of compression index. ASTM Geotech. Test. J. **4**, (1981).
10. O. Rendon-Herrero, Universal compression index equation. J. Geotech. Eng. Div. **106**, 1179–1200 (1980).
11. Y. Nishida, A brief note on compression index of soil. J. Soil Mech. Found. Div. **82**, 1027-1–1027-14 (1956).
12. B.K. Hough, Basic soils engineering (Ronald Press, New York, 1957).
13. A. Al-Khafaji, O. Andersland, Equations for compression index approximation. J. Geotech. Eng. **118**, 148–153 (1992). https://doi.org/10.1061/(ASCE)0733-9410(1992)118:1(148)
14. C. Wroth, D. Wood, The correlation of index properties with some basic engineering properties of soils. Can. Geotech. J. **15**, 137–145 (1978).
15. G. Spagnoli, S. Shimobe, Statistical analysis of some correlations between compression index and Atterberg limits. Environ. Earth Sci. **79**, 532 (2020).
16. J.M. Duncan, Factors of safety and reliability in geotechnical engineering. J. Geotech. Geoenviron. Eng. **126**, 307–316 (2000).
17. S. Onyejekwe, X. Kang, L. Ge, Assessment of empirical equations for the compression index of fine-grained soils in Missouri. Bull. Eng. Geol. Environ. **74**, 705–716 (2015).
18. M. Ozer, N.S. Isik, M. Orhan, Statistical and neural network assessment of the compression index of clay-bearing soils. Bull. Eng. Geol. Environ. **67**, 537–545 (2008).
19. G.L. Yoon, B.T. Kim, Regression analysis of compression index for Kwangyang marine clay. KSCE J. Civ. Eng. **10**, 415–418 (2006).
20. E. Díaz, G. Spagnoli, A super-learner machine learning model for a global prediction of compression index in clays. Appl. Clay Sci. **249**, 107239 (2024). https://doi.org/10.1016/j.clay.2023.107239
21. A.N. Schofield, P. Wroth, Critical state soil mechanics (McGraw-Hill, London, 1968).
22. P.W. Mayne, Cam-clay predictions of undrained strength. J. Geotech. Eng. Div. **106**, 1219–1242 (1980).
23. F. Moran, M. Proctor, Mueser, N.Y. Rutledge, Study of deep soil stabilization by vertical sand drains (Bureau of Yards and Docks, Department of the Navy, 1958).
24. T. Nagaraj, B. Srinivasa Murthy, A critical reappraisal of compression index equations. Géotechnique **36**, 27–32 (1986).