Big data opportunity in geotechnical engineering: A case study on clay property prediction

Wu Stephen　データ科学研究系　准教授

【Background: prediction of soil properties】

City planning, including construction of infrastructures and buildings, is completely dependent on the soil structure at the city location. Predicting the soil properties is an essential task, yet remains to be very challenging because of the complicated interaction between different factors, such as historical climate change or human intervention. These factors often change drastically from location to location, and the spatial variation can be large even within a single city. On the other hand, state-of-the-art tests to measure the soil properties of interest are typically very expensive, creating a practical tradeoff between financial concern and prediction accuracy. There are many soil property data around the world, but the data is very sparse at a single site. Engineers have been relying on expert knowledge to compensate the highly uncertain prediction due to limited data when making important decisions based on the prediction property values.

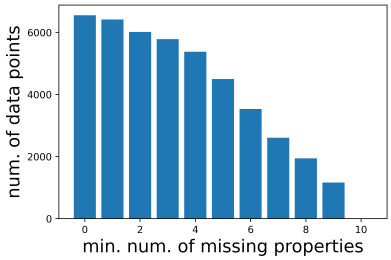
Starting from 2012, Ching and Phoon have devoted their research into building **the first global database for clay properties**. In their latest publication in 2014, their database, called CLAY/10/7490, has grown from the original 345 data points covering only 37 sites to a total of 7490 data points covering more than 600 sites taken from 251 studies. Since then, researchers in geotechnical engineering began to consider a data-driven approach to understand soil properties. However, the use of modern data science tools in this field is still at its infancy. In this study, I will extract some interesting statistics out of the CLAY/10/7490 database based on 10 properties:
three index properties – Liquid Limit (LL), Liquidity Index (LI), and Plasticity Index (PI);
four stress and strength parameters – normalized vertical effective stress (sv/Pa), normalized preconsolidation stress (sp/Pa), normalized undrained shear strength (su/sv), and sensitivity (St);
three cone penetration test parameters – pore pressure ratio (Bq), normalized cone tip resistance (qt1), and normalized effective cone tip resistance (qtu).

In particular, prediction of su/sv based on other properties is the ultimate goal in this case study, as it is the most difficult and expensive property to be predicted in practice.

【Basic statistics of CLAY/10/7490 database】

Database of soil properties often contains many missing data due to the lack of resources to make complete measurements for all properties along different depth of a site. Histogram on the right below shows the distribution of data points as a function of the number of missing properties and table on the left shows the number of data points / the number of sites with at least three common data points shared between a pair of clay properties. The latter number is particular important because we will model the data using normal distribution and the relationships between different clay properties in each site are represented using the coefficient of correlation.

	LL	PI	LI	sv/Pa	sp/Pa	su/sv	St	Bq	qt1	qtu
LL	4451 / 384	-	-	-	-	-	-	-	-	-
PI	4447 / 384	5052 / 420	-	-	-	-	-	-	-	-
LI	3720 / 320	3733 / 322	3923 / 327	-	-	-	-	-	-	-
sv/Pa	3053 / 271	3173 / 281	2633 / 221	3573 / 303	-	-	-	-	-	-
sp/Pa	1831 / 207	1939 / 217	1477 / 165	2104 / 228	2135 / 232	-	-	-	-	-
su/sv	2812 / 271	3264 / 301	2366 / 215	2992 / 280	1863 / 221	4203 / 331	-	-	-	-
St	1618 / 156	1636 / 159	1653 / 148	1025 / 101	651 / 81	978 / 100	1943 / 172	-	-	-
Bq	796 / 69	796 / 69	774 / 68	938 / 76	605 / 62	690 / 66	286 / 24	989 / 80	-	-
qt1	873 / 75	873 / 75	851 / 74	1023 / 82	653 / 67	769 / 72	319 / 26	938 / 76	1023 / 82	-
qtu	796 / 69	796 / 69	774 / 68	938 / 76	605 / 62	690 / 66	286 / 24	938 / 76	938 / 76	938 / 76



【Hidden data pattern in CLAY/10/7490 database】

The pairplot on the left below shows all the property data across different sites. We call this the **Global perspective**. The correlations between ln(su/sv) and other other properties are very weak. However, if we separate the data from the same site and observe the correlations within each site, the distribution of correlations among the sites illustrated that ln(su/sv) is perfectly predictable from other properties for some of the sites. For example, the unclear correlation between ln(su/sv) and ln(sv/Pa) is contaminated by just a small number of sites with opposite correlations. We call this the **Site-specific perspective**. This result demonstrates that a hierarchical probabilistic model shall be used to separate the intra- and inter-site uncertainties in order to see the hidden data pattern covered up by the large noise in the database. The bottom right correlation graphs show a clear correlation relationship between clay properties after we shift from Global perspective to Site-specific perspective. Our group is the first to propose such kind of analysis for soil property prediction and we expect this simple well-known idea in statistics will open up many new opportunities to break the existing challenges in geotechnical engineering, leading to new theories for understanding the complicated soil mechanics in the macro scale.