Practical approach for soil characterization with multivariate analysis

S. Peña. Corresponding / primary author
Acciona Ingeniería, Alcobendas, Spain, Santiago.pena.fernandez@acciona.com

A. Deu., M. Devincenzi
Igeotest, Figueres, Spain, amadeu@igeotest.com, marcelo@igeotest.com,

ABSTRACT: Within the framework of the development of the Port of Barcelona, the Port Authority of Barcelona designed and performed a geotechnical survey consisting on 103 CPTU, 85 boreholes and 11 DMTs from a set of jack-up platforms. This signifies over a million of records only provided by the CPTU. Many analyses were performed with all the information available. Among them, some unsupervised machine learning techniques such as a basic k-means clustering and a more advanced hierarchical clustering showed a very efficient and clear method to objectively characterize the soil behavior and display a visual-spatial distribution. This also allowed the geotechnical team to focus on some unexpectedly soft behavior of the soil at certain depths and specific locations. This paper also shows some of the arguments and conclusions regarding this specific project.

Keywords: CPTU, Characterization, clustering, Machine learning

1. Introduction

Port Authority of Barcelona, has promoted an intensive geotechnical survey in the port of Barcelona. The survey is intended to fully characterize seabed and marine soils for future projects for the port expansion.

Field work was carried out during 217 days. 85 boreholes with continuous sampling, SPT tests and thin wall undisturbed samples were taken together with 103 CPTU and 11 DMT tests. Average investigation depth was around 40m each.

Some positions where executed duplicating the tests, doing both a CPTu and DMT in adjacent locations, or even triple (doing CPTU, DMT and a borehole as well), allowing direct comparison.

Finally, an extensive laboratory tests program was conducted with the undirstubed samples taken inside the boreholes.

The investigation spanned along roughly 2500m parallel to the actual West quay, as shown in Figure 1

Figure 1. Sketch of the survey with CPTs (red), boreholes (grey) and DMTs (green)

The operations were performed by experienced technicians using a 20t CPT rig and hydraulic drilling rigs installed on a jack-up platform deck. Equipment and procedures were in accordance with ASTM D 5778. Some DMT tests following ASTM D 6635 were also made with the same pushing rig and crew.

2. Geological environment

Several geotechnical surveys using geotechnical in situ tests (e.g. CPT and DMT) have been performed along the years at the Barcelona Port and the surrounding infrastructures, where Llobregat Delta quaternary sediments extend, south of the city of Barcelona (e.g.[1] [2]).

General location of Llobregat Delta and geological cross-section showing the main lithological units are presented in Figure 2 and Figure 3, respectively.

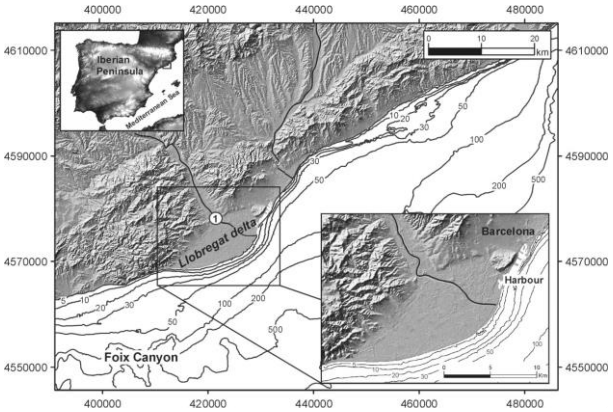

Figure 2. Llobregat delta location and Barcelona harbor.

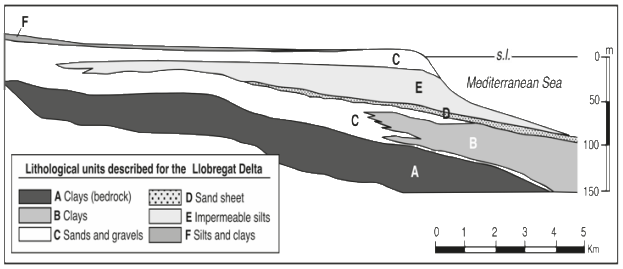

Figure 3. General Longitudinal geological NW-SE cross-section, [1] and [2]. Modified from [3], [4] and [5].

The general architecture of the delta edifice consists of five main units which, from base to top, are:

1. a lower unit of Pliocene blue clays and shales with fragments of shells (bedrock),
2. fluvial gravels that form the "lower aquifer", probably younger than 18.000-15.000 y B.P.
3. pro-delta deposits made of fine grained soils (mainly clayey silts),
4. the "upper aquifer" unit made of gravels, sands and some silts from delta front and delta plain environments and
5. the superficial level with flood plain fine sands, silts and clays and marsh clays ([3], [4]).

3. Machine learning

The last technical committee created by the International Society of Soil Mechanics and Geotechnical Engineer (ISSMGE) is the TC309 focused on disseminating machine learning techniques to solve problems relevant to soils mechanics and geotechnical engineering.

Machine learning techniques are a set of tools and algorithms that transform data sets into actionable knowledge. They can be applied almost in every field of study some way or another. They strengthen engineer understanding boosting their knowledge in an objective and easily understandable manner.

Machine learning techniques take advantage of the huge available data that modern equipment allow to record, in addition with the computing power and statistical methods, resulting in an efficient, reproducible, straight and unbiased approach to analyze and interpret information based on the previous understanding of the problem treated. This is, since this kind of techniques have to be conducted for some purpose and the origin and representativeness of the input data has a paramount impact none of them can be conducted with prior specific understanding. It is worth quoting for this purpose Bret Lantz "*At its core, machine learning is simply a tool that assists us in making sense of the world's complex data. Like any tool, it can be used for good or evil. Machine learning may lead to problems when it is applied so broadly or callously that humans are treated as lab rats, automata, or mindless consumers. A process that may seem harmless may lead to unintended consequences when automated by an emotionless computer"[6]* That is to say, machine learning is not, in any way, a substitute of human brain.

In many occasions geotechnical engineers have to deal with very few information. Even though this methods can be applied within any data set size, they enhances its real potential with large data sets.

An astonishing amount of techniques and approaches can be used for very different purposes: liquefaction, pile capacity, settlements, geomaterial properties, etc… (some of them can be consulted in TC304 webpage) [7].

Machine learning usually is divided in two main groups: supervised or unsupervised learning (sometimes it is also included those called metha-learning algorithms).

Supervised learning is referred to the process of creating a predictive model based on a set of information that, somehow, provides the designer the tools or the information to assess how well the model fits to reality. This models are, hence, predictive.

Within unsupervised learning there is no known output or no tools to assess how well or bad the model behaves. Data is not labeled and, as a consequence, the models developed descriptive (in opposition to the predictive models developed with supervised learning). So the goal of this techniques is analyze the data and extract the information from it.

Many techniques are included in the unsupervised learning. Among them, clustering.

4. Clustering

Clustering includes a set of algorithms intended to create "clusters" or groups of data related or that are similar in some sense. Needless to say how convenient this is for site characterization and as a support for geological mapping (more specific techniques for soil profiling can be used among supervised learning algorithms that are beyond the scope of this paper).

Clustering are a set of non-statistical techniques (in terms of inference, and if not cross-validation or testing is performed) and, probably, as a consequence, easy to visualize and interpret.

The typical steps followed in a cluster analysis are described in [8] are:

1. Clustering Elements .Select the entities to be clustered
2. Clustering variables. Select the variables containing enough information to correctly clusterizing
3. Variable standardization. To make variables comparable
4. Measure of association. Selecting a similarity or dissimilarity measure.
5. Clustering method. The core of the process would be selecting the approapiate clustering method.
6. Number of clusters. Selecting the number of clusters can be difficult if no a priori information exists regarding the number of clusters in the data.
7. Interpretation, testing and replication. In the present paper no hypothesis testing or validation is described to test the validation. However, interpretation and judgment shall be applied within the applied discipline involved in clusteing. In this case, geotechnics. Replication or replicability

must also be guartanteed as part of any scientific method.

This paper focuses on a very specific, basic and easily approachable technique of partitioning clustering: k-means. So it will try to display its benefits in terms of cost-effectiveness and reliability for this particular project.

4.1. Distance measures

As aforementioned, clustering consists in forming groups of data so that the pairwise dissimilarities among the values in each cluster tend to be smaller than those between different clusters. This requires, logically, a definition of similarity, or dissimilarity, that allows ordering the multidimensional data according this criteria.

The most common way to define how "close" or similar are some objects is in terms of the dissimilarity of their attributes. Assuming a p-dimensional elements:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j\left(x_{ij}, x_{i'j}\right) \quad (1)$$

Most common measures of dissimilarity or "distance" are:

1. Euclidean distance

$$d(x_{ij}, x_{i'j}) = \left(x_{ij} - x_{i'j}\right)^2 \quad (2)$$

2. Manhattan distance

$$d(x_{ij}, x_{i'j}) = \left|x_{ij} - x_{i'j}\right| \quad (3)$$

3. Correlation based distances

   a. Pearson Correlation distance (which is a particular case of Minkowski's distance)

$$d_{cor}(x_{ij}, x_{i'j}) = 1 - \frac{\sum_{j=1}^{p}(x_{ij}-\bar{x_i})(x_{i'j}-\bar{x_{i'}})}{\sqrt{\sum_{j=1}^{p}(x_{ij}-\bar{x_i})^2}\sqrt{\sum_{j=1}^{p}(x_{i'j}-\bar{x_{i'}})^2}} \quad (4)$$

   b. Eisen cosine correlation distance [9]

$$d_{eisen}(x_{ij}, x_{i'j}) = 1 - \frac{\sum_{j=1}^{p}(x_{ij})(x_{i'j})}{\sqrt{\sum_{j=1}^{p}(x_{ij})^2}\sqrt{\sum_{j=1}^{p}(x_{i'j})^2}} \quad (5)$$

   c. Spearman correlation distance (obtained by applying the Pearson coefficient to rank-transformed data

   d. Kendall Correlation distance (which counts the number of pairwise disagreements between two ranking lists)

The selection of distance measurement could be critical for an appropriate performance of the model. Depending on the approach, one or another could be more appropriate. Euclidean distance is intuitive and easy to interpret. Pearson's correlation is very commonly used because it measures the degree of linear relationship between variables (common assumption in regular correlations for geotechnical approaches). Eisen's correlation is the cosine of the angle between two vectors connecting the origin to the i^{th} and j^{th} p-dimensional observation respectively. A similar interpretation as Pearson's, although the vectors start from the 'mean' of the p-dimensional observation.

There is a priori no preferential choice in terms of distances. Correlation-based distances assume variables are "close" if they have a good correlation even though the observed values are far apart one from another. According the author's experience, once the normalized CPTu parameters are used (Qt, Fr and/or Bq or similar), and the variables are standardized, Euclidean distance performs a very good approach and tend to create homogenous groups in terms of geotechnical classification.

It goes without saying that, given the variability in range, all the variables used have to be standardized or put them on the same scale in order to compare them (or correctly assessing any "distance" definition without unjustified weighting any of the components of each element). It is commonly done by subtracting the mean and dividing by the standard deviation of each variable. So all the variables are expressed in terms of the numbers of standard deviations with respect of its mean value.

4.2. K-means clustering

For this particular project, among the different potential approaches of finding clusters in a set of data points, it has been used the K-means algorithm [10]. It is, likely, the most commonly used algorithm for clustering. It is also the base for other more complex and modern clustering techniques. It consists on an iterative updated partition by simultaneously relocating each object to the group to whose mean it was closest and then recalculating the group means [11].

K-means proceeds iteratively minimizing the sum of within cluster variance. Since it is only internal variance what is to be minimized, it is implicitly Euclidean distance what is used. Given variance is the sum of squared deviations from the centroid (arithmetic mean position of all the points) it is also the mean Euclidean distance of all the points to the centroid.

K-means procedure is as follows:

- Chose the number of clusters to create (n)
- Randomly place these initial "n" points in the plane/hyperplane
- Create n clusters with the nearest points to each one of these initial points.
- Calculate the centroid for each cluster
- Iteratively create clusters with the nearest points to each of the centroids created in the previous step.
- Proceed iteratively until no further change in the clusters happen.

This proceed is shown in Figure 4 for two clusters where a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centers μ1 and μ2 are shown by the red and blue crosses, respectively. (b) In the initial E step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster center is nearer. This is

equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centers, shown by the magenta line, they lie on. (c) In the subsequent M step, each cluster center is re-computed to be the mean of the points assigned to the corresponding cluster. (d)–(i) show successive E and M steps through to final convergence of the algorithm

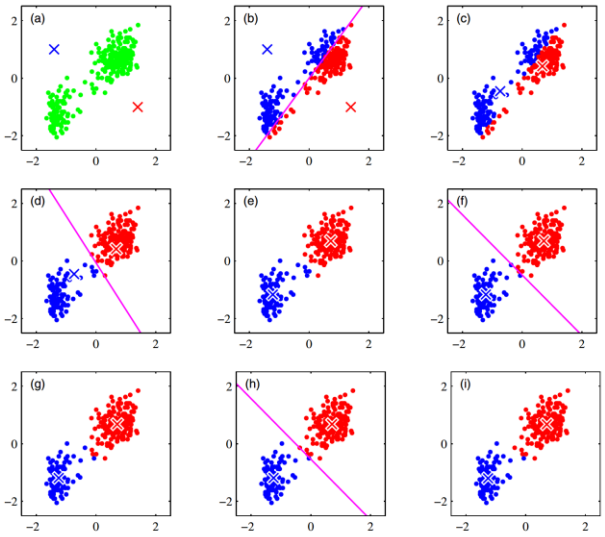

Figure 4. Example extracted from [12]

4.3. Number of clusters to initiate computation

One drawback worth considering about k-means clustering is the fact that it requires specifying the number of clusters to start with. The cluster solution is usually sensitive to the number of clusters selected and the initial selection (location) of cluster centers.

The number of clusters to start with is always a trade-off. The more the number of clusters, the less the sum of within clusters variance. But also the higher the number of groups, the less representative and the less quality the representation has. It is usually solved by plotting the within groups sums of squares vs. the number of clusters extracted Figure 5. It is commonly assumed that the inflexion point represents the optimum balance for the number of clusters decision.

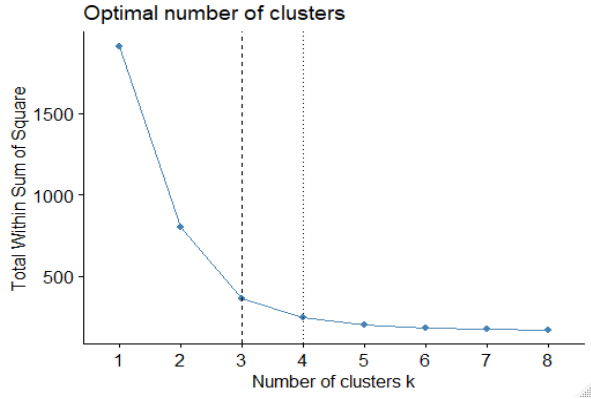

Figure 5. Example of trade-off analysis performed for sieve analysis clustering (using percentage of gravel, sand, silt and clay) for the survey. Optimal options.

5. Barcelona's port survey

As mentioned, Barcelona's port survey included 103 CPTu's with an average length around 40m each. This results, for each of the basic parameters (q_t, f_s and u_2), over 400000 readings.

Given the accuracy and density of information, the clustering itself allows to visualize the geotechnical profile or mapping of the survey with the CPTs alone.

5.1. Numbers of clusters

It has been predefined the number of clusters to be used either for CPTs and laboratory test separately.

5.1.1. CPTs

To estimate the parameters and process data, the team that processed the data assumed a constant unit weight of 18 kN/m^3. Which is, roughly, the average value as estimated in [13] and represented in Figure 6). It was used Normalized Cone Resistance (Qt), Normalized FriCtion Ratio (Fr) and Pore Pressure Ratio (Bq) as they are defined in [14]. The use of non normalized parameters would require somehow including the effective stress as a variable adding computational cost. So using the normalized parameters can be considered a dimension reduction. The output of these other assumptions are not displayed in this paper.

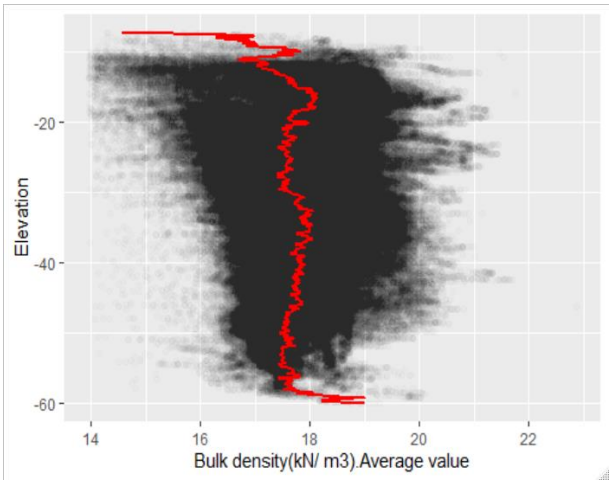

Figure 6. Average bulk density with depth as inferred from the CPTu by [13]

It has been performed several iterations with a 1% (meaning 12000) random sample of the readings and averaged the results of all of them. It is worth commenting that the impact of a small sampling could affect somehow the appropiate number of clusters if there would be a small cluster not represented by the sampling. That would mean that there could be a potential low number of clusters. Visualization and engineering judgement are of utmost importance.

All the outputs resulted in a very similar "shape" to obtain the optimal number of clusters (Figure 7). The output showed that the optimal number would range from 3 to 5.

5.1.2. Sieve analysis

Similarly to CPTs, data a cluster analysis for sieve analysis is displayed in the present paper. Particle size analysis have been chosen because it is usually the laboratory test with the most number of results, so the population is significant enough. It is noted that sieve analysis would not fully represent soil behavior but it seems a reasonable approach for soil characterization at first approach. Additional analysis with regard to this point are not presented in this paper.

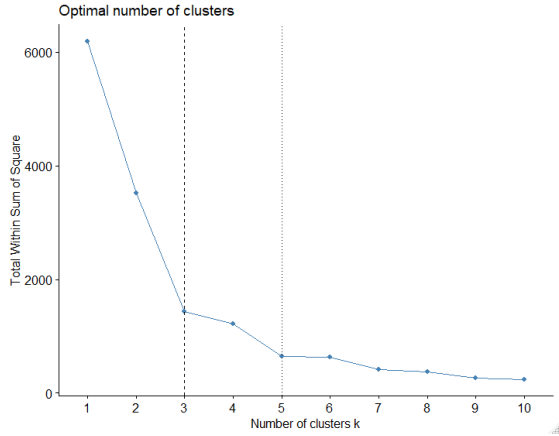

Figure 7. Optimal number of clusters assuming Qt, Fr, and Bq parameters.

Over 638 sieve analysis tests were performed in the survey. 127 sedimentation analysis were also conducted. The following plot shows only cluster analysis performed with the porcentage of gravel, sand and fines (clay+silt). It is used the complete population of tests.

It looks from **Figure 8** that the optimal number of clusters would be 3 or 4.

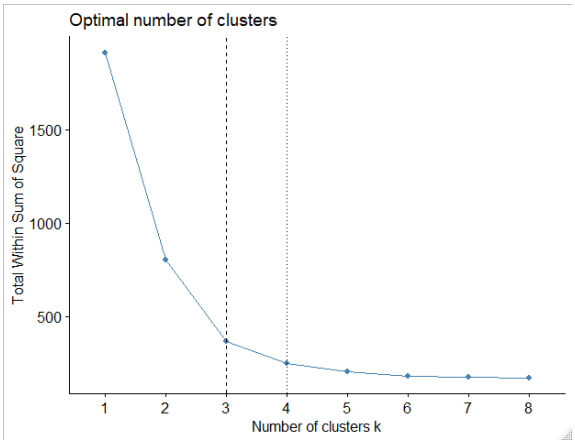

Figure 8. Optimal number of clusters assuming %Gravel, %Sand and % Fines as parameters.

5.2. Cluster results

Several cluster analysis were performed using different parameters and number of clusters. Since the solution could be sensitive to the initial starting partitions, trying several random starting points is often recommended. Cluster analysis is performed using 30 initial random locations for the starting partitions.

5.2.1. CPTUs

The following sections display some of the results obtained for the clustering analysis performed with the normalized CPTu parameters (Qt, Fr and Bq)

- **Five clusters**

The three dimensional model allows to display the geological layering and deposition stages. As it effectively shows the reality (in terms of the CPT output), it also can be seen the interlayered and mixed up fact within the general geological context.

The survey included several long boreholes (60m long) to detect the deep aquifer. All of these boreholes (and also some of the CPTus) reached the aquifer. With these elevations it has been also calculated the regression plane of the aquifer. This plane agrees with the general stratigraphy as it can also be seen in Figure 10 where are represented the five cluster CPT layering, elevation where the aquifer was detected and the regression plane computed with these single points.

As it can be seen, a five groups clusterization results in splitting into two different clusters the SBT=3 soil. In terms of Bq, three out of five groups are essentially drained. The two groups that can be considered with SBT=3 can be joined together, thus resulting 4 groups.

It is worth noticing the similarity with the four clusters approach.

- **Four clusters**

The results of a four cluster level plotted in Robertson's chart are displayed in Figure 12.

Indeed, a reasonably visual approach of the geotechnical profile is displayed, represented on a 3D model, as well as in Figure 13.

When represented in a 2D paper sheet like this, overplotting do not allow to completely visualize the highly interbedded silty- sandy layers found in the geological profile.

This four groups have the basic statitics included in Table 1.

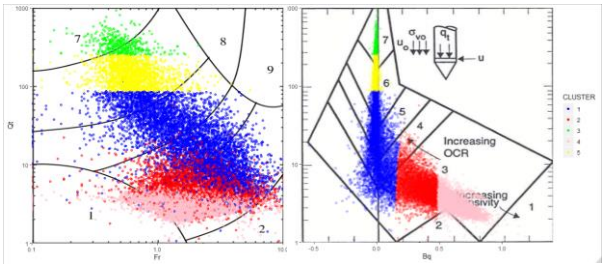

Figure 9. Five Clusters. It is represented a random sample of 5% of the readings in Robertson's charts. (Qt, Fr and Bq as variables)

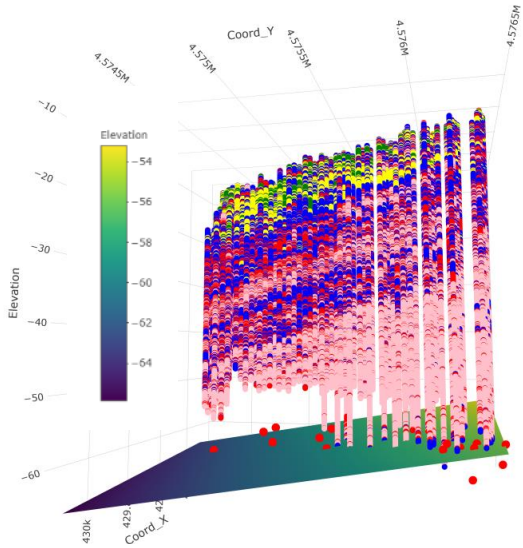

Figure 10. CPTU five Clusters. 3D representation of the profile (Qt, Fr and Bq as variables). Regression plane of the aquifer location.

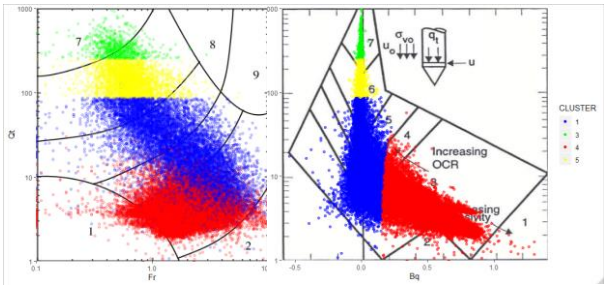

Figure 11. Four Clusters. It is represented a random sample of 5% of the readings in Robertson's charts. (Qt, Fr and Bq as variables)

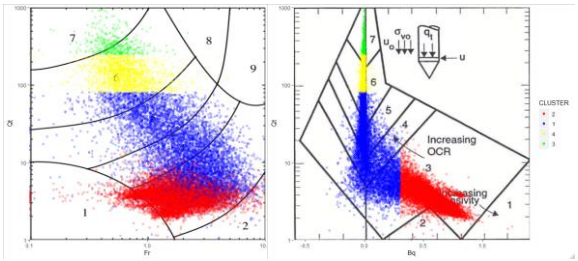

Figure 12. Four Clusters. It is represented a random sample of 1% of the readings in Robertson's charts. (Qt, Fr and Bq as variables)

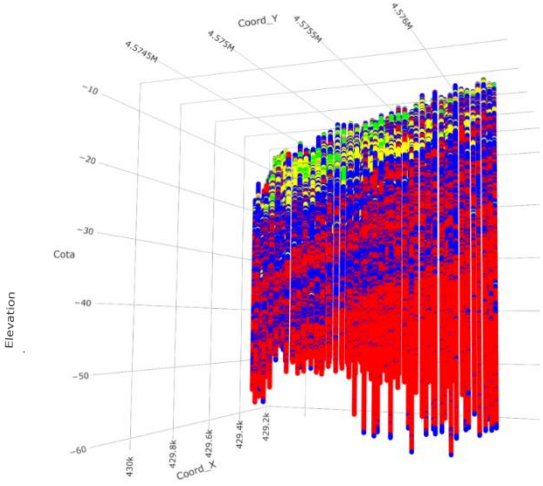

Figure 13. Four Clusters. 3D representation of the profile (Qt, Fr and Bq as variables)

Table 1 .Qt, Fr and Bq statistics for four clusters. Median, mean and first and third quartiles.

CLUSTER 1 (RED) Clay – silty clay		
Qt	*Fr*	*Bq*
1st Qu.: 3.712	1st Qu.: 1.220	1st Qu.: 0.4730
Median : 4.478	Median : 1.670	Median : 0.5680
Mean : 4.673	Mean : 1.986	Mean : 0.5638
3rd Qu.: 5.270	3rd Qu.: 2.400	3rd Qu.: 0.6440
CLUSTER 2 (BLUE) Silt and sand mixtures		
Qt	*Fr*	*Bq*
1st Qu.: 7.879	1st Qu.: 1.200	1st Qu.:-0.01400
Median : 14.997	Median : 2.040	Median : 0.01600
Mean : 24.694	Mean : 2.712	Mean : 0.06025
3rd Qu.: 34.626	3rd Qu.: 3.220	3rd Qu.: 0.13900
CLUSTER 3 (YELLOW) Sand- Silty sand		
Qt	*Fr*	*Bq*
1st Qu.:124.39	1st Qu.: 0.5000	1st Qu.: -0.01100
Median :160.43	Median : 0.6300	Median : -0.00500
Mean :168.92	Mean : 0.7115	Mean : -0.00845
3rd Qu.:205.79	3rd Qu.: 0.8200	3rd Qu.: -0.00100
CLUSTER 4 (GREEN) Gravelly sand – sand		
Qt	*Fr*	*Bq*
1st Qu.: 330.9	1st Qu.: 0.4100 1	st Qu.:-0.008000
Median : 379.6	Median : 0.5100	edian :-0.003000
Mean : 418.5	Mean : 0.5409	ean :-0.004451
3rd Qu.: 453.5	3rd Qu.: 0.6300	rd Qu.:-0.001000

- **Four clusters with Ic**
  - Ic and Bq

An interesting approach is based on Ic (as defined by [15]) and Bq clustering. This could be also understood as an additional dimension reduction. Ic already carries the information that associates Qt-Fr to the soil behavior while Bq somehow provides additional information related with the undrained/drained behavior. It results in an optimal 3 or 4 clustering Figure 14.

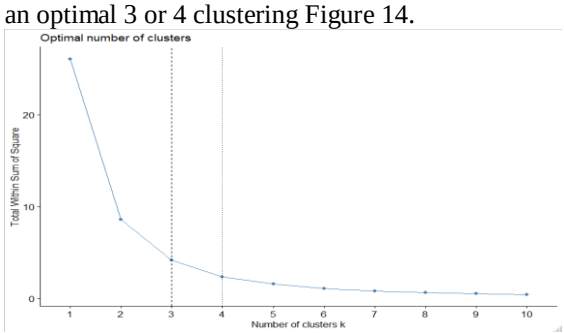

Figure 14. Optimal number of clusters assuming Ic and Bq as variables.

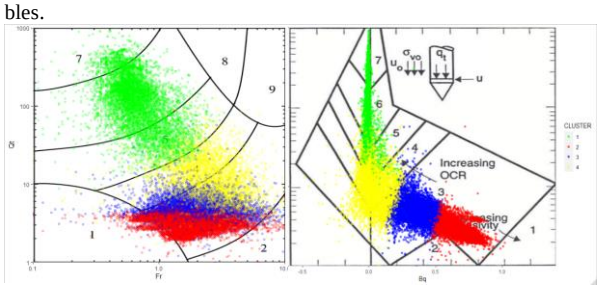

Figure 15. Four Clusters. It is represented a random sample of 5% of the readings in Robertson's charts. (Ic and Bq variables)

As it is displayed in Robertson's chart it results in a highly over plotted area for SBT=3 in the Qt-Fr chart. It allows to differentiate two behaviors based in the Pore Pressure Ratio. However, it does not seem to make much difference in terms of soil behavior characterization than previous approaches.

- Ic and normalized excess pore pressure $(\Delta u_2 / \sigma'_{v0})$

Schneider et al [16] proposed the normalized excess pore pressure a better parameter for the normalized pore pressure. This chart focuses in differentiating fine-grained soils with enough pore pressures and small cone resistances; and allows to distinguish, based on the normalized excess pore pressure between sensitive clays, silts and "low Ir" clays and regular clays. Soil profile in Barcelona port is highly interlayered. Given Robertson's chart do not differentiate visually, based on the Bq parameter, different clay behaviors, it was considered useful using this chart.

The following pictures show the 3 and 4 groups clusterization.

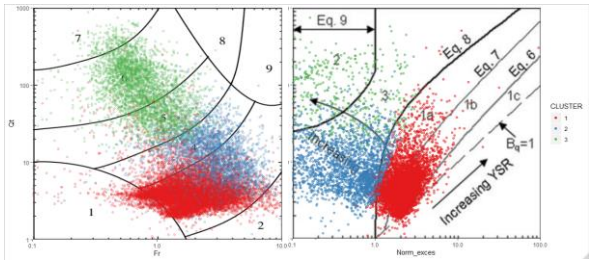

Figure 16. Three Clusters. It is represented a random sample of 5% of the readings in Robertson's charts. (Ic and $\Delta u_2 / \sigma'_{v0}$ variables)

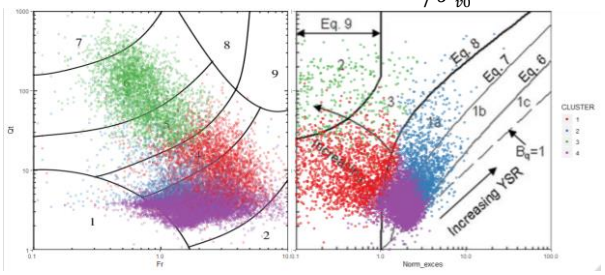

Figure 17. Four Clusters. It is represented a random sample of 5% of the readings in Robertson's charts. (Ic and $\Delta u_2 / \sigma'_{v0}$ variables)

Although this model tends to clearly differenciate essentially drained, transitional and undrained soils, it does not differenciate silts and low Ir clays, from undrained clays. Actually, clustering using Bq as variable (along with Qt and Fr), display a better distinction between 1a (silts or low plasticity clays) and 1b (clays), as shown in Figure 18.

Other approches have been made but not included in this paper.

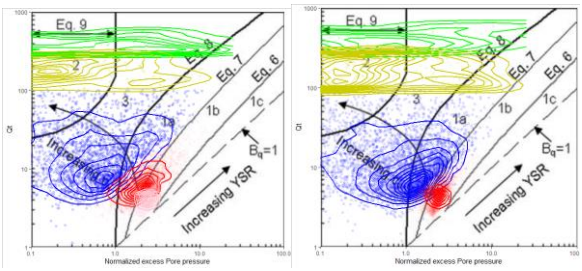

Figure 18. Five (in the left) and four (in the right) clusters (Qt, Fr and Bq as variables). Plotted results over Schneider's chart.

5.2.2. Sieve analysis

Similar analysis were made with laboratory tests, using granulometry. This is, very commonly, the most numerous laboratory test in typical surveys, and that's why they suit very well this type of analysis.

The computational cost of this clustering is not significant, since usually the number of samples do not exceed some tenths. Current survey performed 638 sieve analysis tests.

Unfortunately the randomness, hence the statistical representativeness, in the selection of the samples tested can not be guaranteed. In fact, the sampling can be sometimes biased due to the laboratory technician perspective of the importance of each sample, or the easiness for carving, etc… Even though they could be not considered a random sampling within the soil "population", they can be used to get the big picture of the geology.

- **Three clusters**

In terms of sieve analysis, the optimal number of clusters is estimated in three (see Figure 5) with the basic statistics included in Table 2.

Table 2 **.**Sieve analysis statistics for three clusters. Median, mean and first and third quartiles.

CLUSTER 1		
Gravel	*Sand*	*Fines*
1st Qu.:25.20	1st Qu.:33.25	1st Qu.:15.95
Median :28.60	Median :36.50	Median :28.10
Mean :34.78	Mean :38.46	Mean :26.75
3rd Qu.:38.80	3rd Qu.:44.65	3rd Qu.:38.80
CLUSTER 2		
Gravel	*Sand*	*Fines*
1st Qu.:0.00000	1st Qu.:1.150	1st Qu.:86.65
Median :0.00000	Median :4.000	Median :96.00
Mean :0.05508	Mean :8.521	Mean :91.40
3rd Qu.:0.00000	3rd Qu.:13.25	3rd Qu.:98.80
CLUSTER 3		
Gravel	*Sand*	*Fines*
1st Qu.:0.000	1st Qu.:50.40	1st Qu.:25.55
Median :0.000	Median :61.85	Median :36.25
Mean :1.474	Mean :62.10	Mean :36.43
3rd Qu.: 0.325	3rd Qu.:74.12	3rd Qu.:48.77

- **Four clusters**

Table 3 **.**Sieve analysis statistics for four clusters. Median, mean and first and third quartiles.

CLUSTER 1 (n=101)		
Gravel	*Sand*	*Fines*
1st Qu.:0.0000	1st Qu.:27.40	1st Qu.:54.70
Median :0.0000	Median :34.10	Median :65.40
Mean :0.2386	Mean :35.58	Mean :63.99
3rd Qu.:0.0000	3rd Qu.:45.30	3rd Qu.:71.60
CLUSTER 2 (n=81)		
Gravel	*Sand*	*Fines*
1st Qu.: 0.000	1st Qu.:61.80	1st Qu.:21.00
Median : 0.100	Median :68.00	Median :31.80
Mean : 1.927	Mean :69.74	Mean :28.33
3rd Qu.: 0.700	3rd Qu.:77.90	3rd Qu.:36.60
CLUSTER 3 (401)		
Gravel	*Sand*	*Fines*
1st Qu.:0.00000	1st Qu.: 1.000	1st Qu.:92.1
Median :0.00000	Median : 2.900	Median :97.0
Mean :0.05536	Mean : 5.374	Mean :94.6
3rd Qu.:0.00000	3rd Qu.: 7.900	3rd Qu.:99.0
CLUSTER 4 (n=11)		
Gravel	*Sand*	*Fines*
1st Qu.:25.20	1st Qu.:33.25	1st Qu.:15.95
Median :28.60	Median :36.50	Median :28.10
Mean :34.78	Mean :38.46	Mean :26.75
3rd Qu.:38.80	3rd Qu.:44.65	3rd Qu.:38.80

5.2.3. Comparison

Both sieve analysis (based on gravel, sand and fines percentages) and CPTu optimal number of clusters are three or four. Above this number of clusters, sum of variances decreases significantly more slowly.

It is worth remembering soil behavior type do not try to match sieve analysis or USCS classification rather than a representation of the behavior during the penetration. However, in this case, the CPTU three cluster's shows there is one cluster covering a wide range of soil behaviors Figure 19 (eventhough the 2D density plot apparently shows it contains two different areas of concetration), and it is also what sieve analysis shows in Table 2 for cluster 2.

Similarly, the four clusters output shows a clustering covering a range significantly wide (Figure 20).

Eventhough it is not assessed quantitatively (and it would imply at certain stage a hypothesis testing and certain rules to compare SBT and sieve analysis to compare them) here, soil behavior and sieve analysis roughly matches in terms of soil characterization. CPTu, likely, reflects better the existence of a large ranges of transitional soils coming from a very interlayered structure.

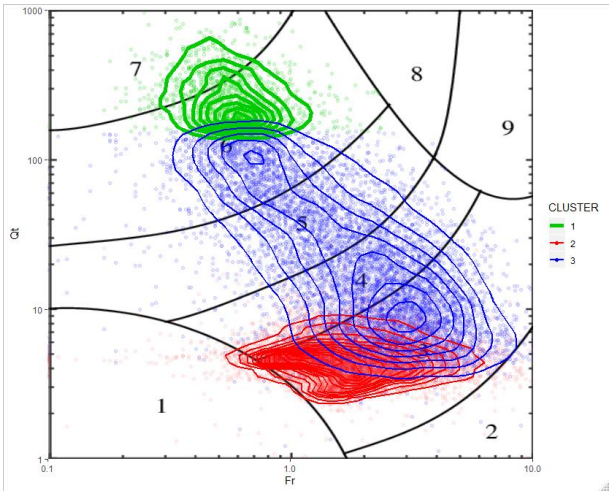

Figure 19. Three Clusters analysis. . It is represented a random sample of 5% of the readings in Robertson's charts with density contours. (Qt, Fr and Bq as variables)

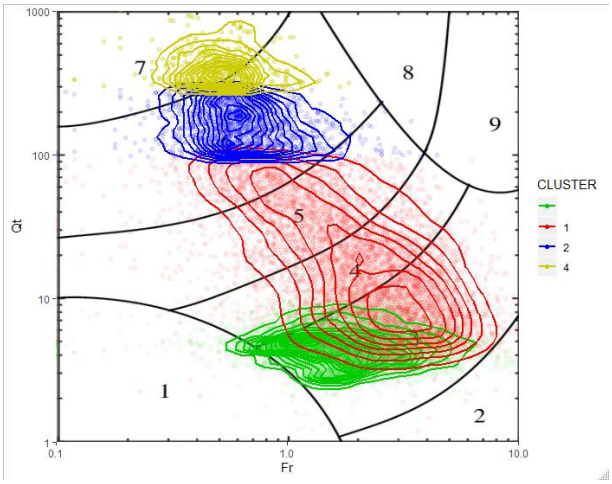

Figure 20. Four Clusters analysis. . It is represented a random sample of 5% of the readings in Robertson's charts with density contours. (Qt, Fr and Bq as variables)

6. Hierarchical clustering

As aforementioned, k-means clustering can be sensitive to the number and initial location of the cluster centroids. This issue can be bypassed, if necessary, by performing a hierarchical clustering. Hierarchical clustering is a procedure that does not require predefining the number of clusters because it is defined in a scalable way.

Hierarchical clustering are either divisive or agglomerative .If divisive, the algorithm assumes all the points form one unique cluster and proceeds progressively dividing the most heterogeneous cluster (in the first step is the only cluster) into two clusters. If agglomerative, it starts assuming each element as a cluster itself and, progressively, creates clusters by merging the closest clusters (or the individual elements in the first step). It ends up with all the elements merged in a unique cluster. This process can be represented with a dendogram Figure 21. Agglomerative is, likely, the most common hierarchical clustering process and it requires a dissimilarity matrix, hence pre-defining a "distance" as a measure of similarity.

In order to improve k-means results, it can be performed a so-called hierarchical k-means clustering which acts as a hybrid model using the centers of the clusters of an agglomerative hierarchical process followed by a k-means cluster analysis.

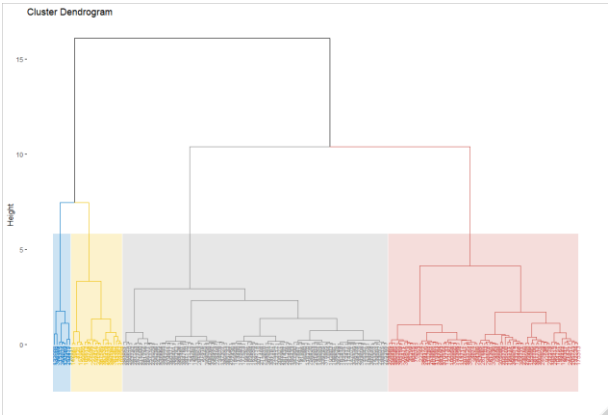

Figure 21. Cluster endogram representing the agglomerative analysis of the 0.5% of the CPTu readings. Split at 4 level clusters.

In this project, this method proved to effectively discriminate the softest areas when splitting the readings into 6 clusters. It is located in the northern area and it dips north-east with similar slopes to the aquifer.

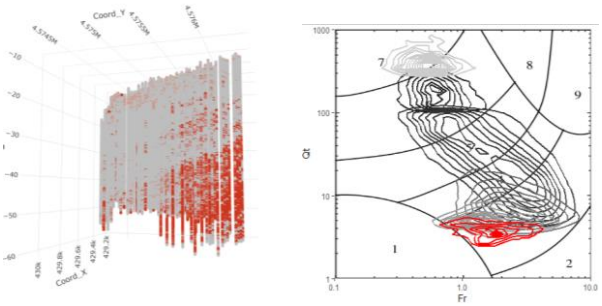

Figure 22. Localized softer areas with Qt of 3.74 ± 1.08

7. Conclusions

Geotechnical in situ tests, CPTU and DMT in particular, have proved particularly useful for stratigraphic and geotechnical characterization of the sediments of the Llobregat River delta.

Several machine learning algorithms are very suitable for the geotechnical engineer every day job escalating engineer's capacities and judgement. Among them k-means clustering is considered to be a very simple and useful approach for soil characterization as starting point.

K-means clustering is a very basic unsupervised machine learning technique for finding groups of affinity in a data set. However, it is of the utmost importance understanding clustering only group data with similar characteristics not specific soil behaviors. Judgment and specific background is required to assess its output.

Most of the benefits come from the cost-effectiveness of this kind of methods (it is almost an automatic and inmediate process) promoting and boosting enginner work.

For this specific project, clustering with basic normalized parameters results in very visual and reasonable approach for soil classification.

Further analysis were performed in this project based on this technique and followed by prototype methods and other supervized tehcniques using Schneider and Robertson's charts that will be presented in the near future.

Acknowledgment

The authors would like to acknowledge Barcelona Port Authority for its commitment with soil investigation and its perception and vision of the importance of geotechnics within the scope of civil engineering. Without it, this studies and those to come would not have been possible.

It would not have been possible either without the cooperation and guidance of Barcelona politechnic university geotechnical engineering and geo-science department, with special thanks to proffesor Antonio Gens and Daniel Tarragó.

References

[1] Devincenzi, M. Colas, S., Casamor, J.L., Canals, M., Falivene, O. & Busquets, P. 2004. "High resolution stratigraphic and sedimentological analysis of Llobregat delta nearby Barcelona from CPT & CPTU tests". Proc. 2 nd International Conference on Site Characterization ISC'2, Porto. Vol 1. Millpress, Rotterdam.

[2] Lafuerza, S., Canals, M., Casamor, J.L., Devincenzi, M.J.. 2005. "Characterization of deltaic sediment bodies based on in situ CPT/CPTU profiles: A case study on the Llobregat delta plain, Barcelona, Spain." Marine Geology 222–223 (2005) 497 – 510.

[3] Marquès, M.A., 1974 "Las formaciones cuaternarias del Delta del Llobregat". Tesis doctoral. Universitat de Barcelona, Barcelona. Pp. 401.

[4] Bayó, A. 1985. "Les aigües Recursos i riscos geològics". Història Natural dels Pa-ïsos Catalans, 3. Enciclopèdia Catalana. Barcelona.

[5] Ventayol, A., 2003. "Caracterización Geotécnica de sedimentos deltaicos mediante piezoconos. Aplicación al margen izquierdo del delta del Llobregat". Ingeniería del Terreno. INGEOTER Vol. 2, C. López Jimeno (ed.): 413-433.

[6] Lantz, B. "Machine Learning with R"

[7] Technical committee web;http://140.112.12.21/issmge/tc304.htm

[8] Milligan, G.W. (1996) Clustering validation: results and implications for applied analyses, in Clustering and Classification (P. Arabie, L. J. Hubert and G. De Soete, eds) 341–375. World Scientific. Singapore

[9] Eisen M.B et al "Cluster analysis and display of genome-wide expression patterns" Proceedings of the National Academy of Sciences of the United States of America. Vol. 95, pp. 14863–14868, December 1998 Genetics. 10.1073/pnas.95.25.14863

[10] Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2), 129–137.

[11] Ball, G. H. and Hall, D. J. (1967) "A clustering technique for summarizing multivariate data". Behavioural Science, 12, 153–155. https://doi.org/10.1002/bs.3830120210

[12] Bishop M.C. "Mixture Models and EM", In: Jordan, M., Kleinberg, J., Schölkopf, B. (eds.) Pattern Recognition and machine learning, Springer, Cambridge, 2006.

[13] Robertson, P.K., and Cabal, K.L, 2010, "Estimating Soil Unit Weight form CPT", 2nd Int. Sym. On Cone Penetration Testing, Huntington Beach, CA, USA.

[14] Robertson, P.K. 1990. "Soil classification using the cone penetration test". Canadian Geotechnical Journal, 27(1): 151–158. doi:10.1139/t90-014.

[15] Robertson, P.K. and Wride, C.E., 1998. Evaluating cyclic liquefaction potential using the cone penetration test. Canadian Geotechnical Journal, Ottawa, 35(3): 442-459.

[16] Schneider, J.A., Randolph, M.F., Mayne, P.W. & Ramsey, N.R. 2008. "Analysis of factors influencing soil classification using normalized piezocone tip resistance and pore pressure parame-

ters". *Journal Geotechnical and Geoenvironmental Engrg*. 134 (11): 1569-1586.
[17] Everitt, B.S., Landau, S., Leese, M., Stahl, D., "Cluster analysis". 5th ed., John Wiley & Sons, Ltd
[18] Murphy, P.K. "Clustering", In: The MIT Press, Machine Learning. A probabilistic Perspective, The MIT Press Cambridge, Massachusetts, London, England.
[19] Lloyd, S. P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory 28(2), 129–137. http://dx.doi.org/10.1109/TIT.1982.1056489
[20] Ball, G. H. and Hall, D. J. (1967) "A clustering technique for summarizing multivariate data". Behavioural Science, 12, 153–155. https://doi.org/10.1002/bs.3830120210