TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA CARACTERIZACIÓN Y PERFILADO DEL TERRENO. APLICACIÓN PRÁCTICA AL CPTU.

Applied machine learning and predictive modeling techniques for soil profiling. Practical CPTU application.

Santiago Peña Fernández[a], Enrique Asanza Izquierdo[a]

[a] Laboratorio de Geotecnia del CEDEX, España.

RESUMEN – Este artículo introduce algunos conceptos e ideas básicas del campo del aprendizaje automático, con los que afrontar la caracterización geotécnica y perfilado del terreno a partir de ensayos de penetración tipo CPTU. Se muestran los resultados prácticos obtenidos para una campaña para un puerto español. En última instancia se pretende mostrar al diseñador algunos algoritmos que permitan ayudar a la elaboración, en los casos que sea posible, de modelos de terreno objetivamente evaluables, repetibles y precisos.

SYNOPSIS – This article addresses some concepts and principles of machine learning applied to soil profiling and geotechnical characterization based on cone penetration testing (CPTU). Some practical outcomes from a survey at a Spanish Port are discussed. Ultimately, the article aims to provide a mathematical approach that assist designers to produce, if possible, a more objectively assessable repeatable and precise soil profile models.

Palabras Clave – Aprendizaje automático, caracterización del terreno, CPTU

Keywords – Machine learning, soil characterization, CPTU.

1 – INTRODUCIÓN

Indiscutiblemente, en las últimas décadas se ha producido un cambio de paradigma en la forma de abordar y gestionar la información debido a la ingente cantidad de registros creados, lo que se ha acuñado como la "era del Big Data". Existe, además, gran disponibilidad de mucha información: así, se tiene constancia de más de un trillón de páginas web en el mundo; y el genoma (consistente en $3.8 \cdot 10^9$ pares de bases) de miríadas de personas ha sido ya secuenciado. Los casos en que la cantidad de información era algo que, apenas hace unos años, ni siquiera se valoraba administrar, son ahora bases de datos estándar.

A la par que la disponibilidad y acopio de datos se ha ido incrementando, diversas técnicas matemáticas de análisis de datos han ido adquiriendo una tremenda popularidad. Estas técnicas se han venido a englobar de forma algo general dentro de los denominados algoritmos de aprendizaje automático ("Machine Learning") (Murphy, 2012). En términos laxos, se puede decir que estos algoritmos proveen de métodos o estrategias automatizadas para el análisis de datos. Es entendible

E-mails: Santiago.pena@cedex.es (S. Peña), Enrique.asanza@cedex.es (E. Asanza)

ORCID: orcid.org/0000-0002-1079-7995, orcid.org/0000-0001-5436-0181

que, bajo esta definición, su aplicabilidad se aprecie de forma difusa. No obstante, este conjunto de métodos permite estudiar la información existente (cualquiera que sea su ámbito) para poder extraer conclusiones de manera objetiva, medible y repetible. En general, su aplicación se sustenta en el uso intensivo de la capacidad de cálculo de los equipos actuales. Algunas de estas técnicas son fácilmente comprensibles, o visualizables, y no requieren ningún conocimiento previo en estadística u otras ramas de las matemáticas. Otras, en cambio, son más complejas y se apoyan en métodos cuya comprensión puede requerir una fuerte base de conocimientos matemáticos, y cuya visualización es imposible en la práctica. Para cualquiera de los dos casos, en ocasiones se pueden construir modelos e intentar obtener los resultados que mejor se ajustan al problema al que se pueda enfrentar el ingeniero en su trabajo diario, incluso sin el conocimiento detallado de los algoritmos, siempre que se sigan unas pautas básicas de validación y fiabilidad, de forma algo similar a como se usa una calculadora sin comprender cómo opera internamente el aparato.

El comité técnico TC309 ("Machine Learning"), de reciente creación, de la ISSMGE Sociedad Internacional de Mecánica de Suelos, tiene por objetivo la diseminación del conocimiento y puesta en práctica de las técnicas de aprendizaje automático para resolver problemas relevantes para la mecánica de suelos e ingeniería geotécnica. A través de la página web del TC304, estrechamente vinculado con el anterior, se encuentran disponibles decenas de ejemplos de aplicación de estas técnicas en una amplia variedad de ámbitos dentro del campo de la geotecnia (http://140.112.12.21/issmge/ML_ref.htm).

Si bien se ha comprobado su utilidad en otras campañas, este artículo sólo muestra los resultados del perfil del terreno partiendo exclusivamente de registros de CPTUs. En general, con cualquiera de los algoritmos aplicados, perfiles geológicos simples resultan en modelos similares y sin más complicaciones que la apropiada interpretación de los ensayos in situ. En el caso de perfiles más complejos, laminados, o con intercalaciones, los modelos son más divergentes. En cualquier caso, los principios aplicados a los resultados de los CPTU se pueden trasladar a cualquier ensayo in situ o de laboratorio o registro que sirva para categorizar un punto del espacio.

Puertos del Estado, ente público empresarial dependiente del Ministerio de Transportes, Movilidad y Agenda Urbana de España, ha mostrado interés en que se impulsen estas técnicas en los trabajos geotécnicos de su competencia, por lo que ha dotado al encargo con el CEDEX una partida para acometer esta investigación. Este artículo es un resumen del avance del trabajo.

2 – APRENDIZAJE AUTOMÁTICO

La denominada inteligencia artificial, o el aprendizaje automático, quizá se encuentre actualmente en una fase de sobredimensionamiento, provocado, tal vez, por una sobreexposición mediática que no favorece la profundización sobre su alcance y eficacia. Los beneficios y versatilidad de sus técnicas son evidentes y, muy probablemente, formarán parte a corto plazo de las herramientas de trabajo habituales en un amplio rango de ocupaciones y actividades.

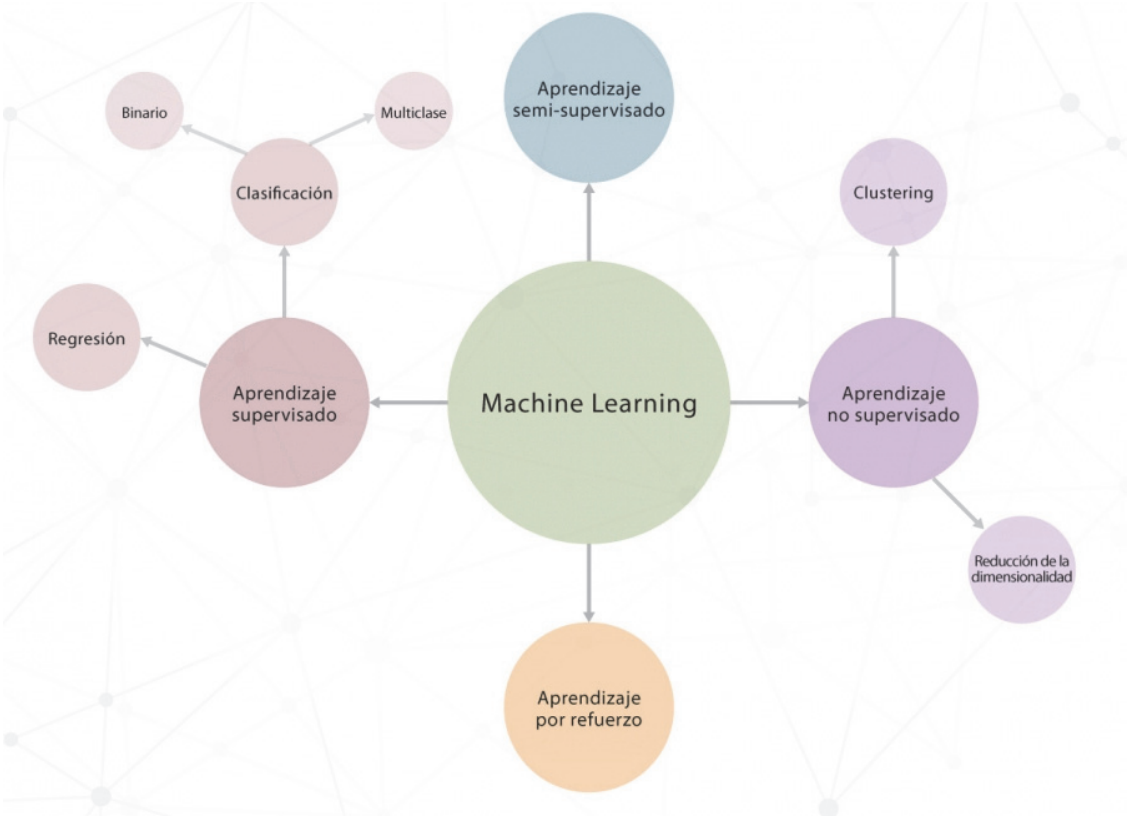
Se trata de un conjunto de procedimientos extremadamente extenso y de estudiada especificidad. Se encuentran, en cualquier caso, lejos de replicar el funcionamiento de la inteligencia humana y, probablemente, como dice Ramón López de Mántaras (fundador y exdirector del instituto de Investigación en Inteligencia Artificial del CSIC), constituyen un ejemplo de lo que Daniel Dennett define como "habilidad sin comprensión" (López Mántaras, 2020).

Existe también en algunos ámbitos cierta confusión o indefinición en la terminología, quizá intencionadamente llamativa: "Inteligencia Artificial", "Aprendizaje automático" (machine learning) o "Aprendizaje profundo" (Deep learning). En primera aproximación, se puede estar de acuerdo en que el aprendizaje automático engloba las técnicas de aprendizaje profundo (que prácticamente hacen referencia, de manera exclusiva, a las redes neuronales), y la inteligencia artificial contiene las técnicas de aprendizaje automático e incluye, de manera genérica, todas aquellas técnicas orientadas a automatizar tareas normalmente realizadas por seres humanos.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

Se utiliza muy ampliamente la terminología original (en inglés) no encontrándose, aún, del todo estandarizada o extendida su traducción al español. A lo largo del artículo se intentará buscar una traducción literal, o aproximada, que, intuitivamente, dé a entender el alcance de lo que la denominación original pretendía o, en su defecto, busque su equivalencia en español (preferiblemente si ya ha sido utilizada en alguna publicación).

Existen diversas clasificaciones de las técnicas de aprendizaje automático (Duda et al (2000), Hastie et al (2008), etc.). Los algoritmos de aprendizaje automático se suelen dividir, en general, en dos grandes grupos: supervisados y no supervisados (se deja al margen el denominado "aprendizaje de refuerzo"). El término "supervisado" hace referencia a la capacidad del modelo de auto-confirmar su capacidad predictiva con un conjunto de datos de prueba reservados a tal efecto; los algoritmos no supervisados hacen referencia a modelos descriptivos sin posibilidad ni pretensión de actuar como predictores. Esto se esquematiza en la Figura 1.



Fuente:*https://datos.gob.es/es/blog/como-aprenden-las-maquinas-machine-learning-y-sus-diferentes-tipos*

Fig. 1 – Tipos de aprendizaje automático.

Los algoritmos de **aprendizaje supervisado**, o **predictivos**, crean modelos para pronosticar valores de una variable respuesta (pongamos m-dimensional) a partir de los valores de una variable predictora (n-dimensional). Estos modelos se crean a partir de un subconjunto de la información disponible, típicamente en torno al 75% de la muestra, a partir del cual se "entrena" el modelo, y se validan, o comprueban, con el subconjunto de la información disponible que no haya sido utilizado para crear el modelo (25% restante). De esta manera se puede decir que el modelo ha "aprendido" de forma "supervisada" puesto que el modelo ha sido generado, probado y evaluado con la información disponible. Esto es, el modelo aporta una respuesta inferida (y^*), y el "supervisor" expone la respuesta correcta (y) comprobando, así, si la respuesta es correcta y, en su caso, el error asociado a la respuesta del modelo (una función de pérdida como, por ejemplo, $L(y^*,y)=(y^*-y)^2$, si el valor es numérico y continuo). En este tipo de problemas también se suele decir que la muestra se encuentra "etiquetada" ("labeled"), en referencia a que lleva asociada la variable objetivo.

El segundo tipo de algoritmos son los denominados de **aprendizaje no supervisado**, o modelos **descriptivos**. Éstos no disponen de un subconjunto de la información para validar o evaluar el modelo (bien sea por voluntad del evaluador o porque la cantidad de información es limitada). Su función es "acopiar" toda la información disponible y observar "patrones interesantes" dentro del conjunto de N observaciones de un k-vector. El objetivo es, entonces, inferir, de alguna manera, las propiedades de los datos sin la "ayuda" de un "supervisor" que evalúe la fiabilidad o el error en cada observación. Se dice, en este sentido, que se trata de muestras "no etiquetadas" ("unlabeled"). Aunque se trata, en general, de un problema más abierto e indefinido que el del aprendizaje supervisado (en tanto que no se predefine qué clase de patrones se buscan o qué medida es apropiada para evaluar el error), y suele abarcar un campo, quizá, menos sugerente, reconocido y con un rango de aplicación menor, se trata de un problema muy recurrente en el trabajo diario de un ingeniero geotécnico y que tiende a resolverse de manera intuitiva y poco estricta.

Existen diversos motivos por los que el aprendizaje no supervisado pudiera ser de interés, varios de los cuales pudieran estar relacionados con el trabajo de caracterización del terreno. Uno de ellos es que puede ser razonable operar de manera inversa, esto es, trabajar de manera no supervisada con grandes cantidades de datos no etiquetados y, posteriormente, crear modelos supervisados. Este es el procedimiento seguido en este artículo. Ciertamente el lector geotécnico intuirá que la aplicación práctica de la mayoría de las técnicas de aprendizaje automático en el trabajo diario, no exige, a priori, la rigurosidad y precisión habituales que los proyectos de inteligencia artificial o los analistas de datos reportan en sus artículos.

3 – CARACTERIZACIÓN GEOTÉCNICA

El proceso de caracterización es quizá uno de los más controvertidos y difusos en Geotecnia. Dicho proceso que, partiendo de un total desconocimiento, ha de discriminar tipos de terreno y asignarles parámetros, lleva aparejado diversas de las fuentes de error e incertidumbres propias de un proceso de análisis. Se resume en la siguiente secuencia:

A. *Acopio de información y datos numéricos*: ensayos in situ y de laboratorio, como registros de sondeos, o cualquier trabajo de campo de que se pueda inferir o extraer algún dato.
B. *Evaluación de la información y datos*: labores de gabinete para su análisis y procesamiento;
C. *Selección del modelo que mejor representa o resume la información*: elaboración de un perfil geológico-geotécnico y la asignación de unos parámetros característicos. Estos procesos se pueden desarrollar en paralelo, secuencial o alternativamente. Este modelo debe ser lo suficientemente simple para su uso práctico y, al mismo tiempo, lo bastante detallado como para no perder capacidad descriptiva.
D. *Estimación de la incertidumbre del modelo y verificación de las asunciones adoptadas*.

Las posibles fuentes de error de estas etapas pueden incluirse dentro de alguno de estos grupos (Nadim, 2007):

- *Incertidumbre por errores u omisiones humanos*. Si se producen, son difíciles de predecir. No obstante, hoy día no puede prescindirse de la labor humana, que desempeña labores correctoras y de supervisión. Todo modelo precisa de esta labor para ser construido de acuerdo con algún criterio establecido por la experiencia, conocimiento o alcance definido por el ser humano. Sólo pueden reducirse los errores humanos si se destinan los trabajos tediosos y repetitivos a algoritmos y se asignan a la persona especialista aquellas tareas que son mejor realizadas por él.
- *Incertidumbre aleatoria*: es la propia de la naturaleza que, por el hecho de estar expuesta a un número muy elevado de agentes actuando sobre ella, muestra una variabilidad natural reflejo del efecto en mayor o menor grado de cada uno de estos agentes. Esta

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

incertidumbre es inherente a cualquier objeto no idealizado. Algunos ejemplos pueden ser la variación de los parámetros del terreno dentro de un mismo estrato, la variación de la altura de ola con el tiempo, etc. El ser humano no puede, a priori, intervenir sobre esta incertidumbre y debe ser aceptada, estudiada y operar en consecuencia.

- *Incertidumbre epistémica:* que es la producida por la falta de conocimiento de la variable estudiada y cuya fuente puede estar localizada en:
    1. *Errores en las mediciones.* Definida como un error (aleatorio en torno al valor medio poblacional) y/o sesgo (constante en un sentido u otro) en el registro de los resultados de los ensayos realizados (precisión del valor bruto del ensayo in situ realizado: SPT, CPT, etc.). En este sentido resulta de importancia tener cierta sensibilidad a la precisión habitual del ensayo realizado y a detectar, aunque resulte, en ocasiones, complejo, si el equipo tiene o no algún sesgo. Ensayos como el SPT, con una significante falta de repetibilidad, incluso en condiciones controladas, debido a su naturaleza dinámica y a la variabilidad de la energía aplicada, puede contener errores difíciles de acotar a priori, pero con un impacto potencialmente significativo en el diseño.
    2. *Errores estadísticos.* Debidos a la falta de información a causa de un tamaño muestral limitado. En la práctica es habitual realizar asunciones poco argumentadas, o sin considerar su efecto sobre la incertidumbre, sobre qué valores representativos adoptar en función del tamaño muestral y tipo de ensayo, y la decisión del valor característico suele estar vinculada a la percepción subjetiva del geotécnico. Las normativas actuales tales como el Eurocódigo 7 (EN 1997-2 (2007)) muestran cierta tendencia a incluir criterios estadísticos en este sentido (no vinculantes) para una evaluación del valor característico razonablemente (asociando un intervalo de confianza concreto) conservador del valor medio. Estas aproximaciones tienen la virtud de poner de relevancia, si procede, la falta de información.
    3. *Errores del modelo aplicado:* provocados por las simplificaciones asumidas al idealizar el problema. Por ejemplo, si se asume que existe una relación biunívoca entre el valor del ángulo de rozamiento y el valor del ensayo de penetración estándar, se está asumiendo un error y/o un sesgo en ocasiones poco valorado.

La secuencia de pasos del proceso de caracterización puede llevar asociada algunos de los errores descritos. En los siguientes apartados se mostrarán algunas propuestas, para reducir, acotar o evaluar algunos de estos errores asociados, fundamentalmente, a los pasos B, C y D que se realizan durante el proceso de caracterización. A continuación se mostrarán algunos resultados aplicados a la campaña de piezoconos mencionada. Por su tamaño muestral, antiguamente, las campañas grandes de CPTU resultaban, en ocasiones, difíciles de manejar u operar, pero contienen una información incomparable, completa y continua para la elaboración de perfiles.

4 – REGISTROS UTILIZADOS EN ESTE ESTUDIO

Para el estudio se han utilizado diversas campañas de piezoconos ejecutadas en ambiente marino. Sólo se mostrarán los resultados de una campaña específica ejecutada desde pontona en el Puerto de Barcelona. En dicha campaña se realizaron (entre otros ensayos y sondeos) 103 CPTUs de unos 40m de profundidad desde el fondo marino. Esto supone, considerando una lectura cada cm, más de 400.000 registros de cada una de las 4 variables medidas por el CPTU: resistencia por punta (q), resistencia por fuste (f_s), presión de poros registrada durante el avance (u_2) y profundidad.

Las propuestas descritas están orientadas a su aplicación en el trabajo diario de un diseñador asumiendo la disponibilidad de un ordenador convencional de gama media-alta. La posibilidad de disponer de equipos de alta gama o computación de alto rendimiento cambia, lógicamente, la

perspectiva de algunos comentarios. Esto último permite el desarrollo de modelos mucho más grandes y sin la necesidad de llevar a cabo estrategias simplificadoras. Dependiendo del algoritmo usado, con ordenadores convencionales, puede resultar impracticable una matriz de 400.000 x 4, por lo que se puede hacer necesario tomar sub-muestras aleatorias de menor tamaño.

Aunque lo descrito en lo que sigue es de aplicación a modelos tridimensionales, por claridad expositiva y de visualización se ha utilizado la proyección sobre una alineación de toda la longitud estudiada. Se ve en planta en la Figura 2 y el perfil utilizado en la Figura 3.

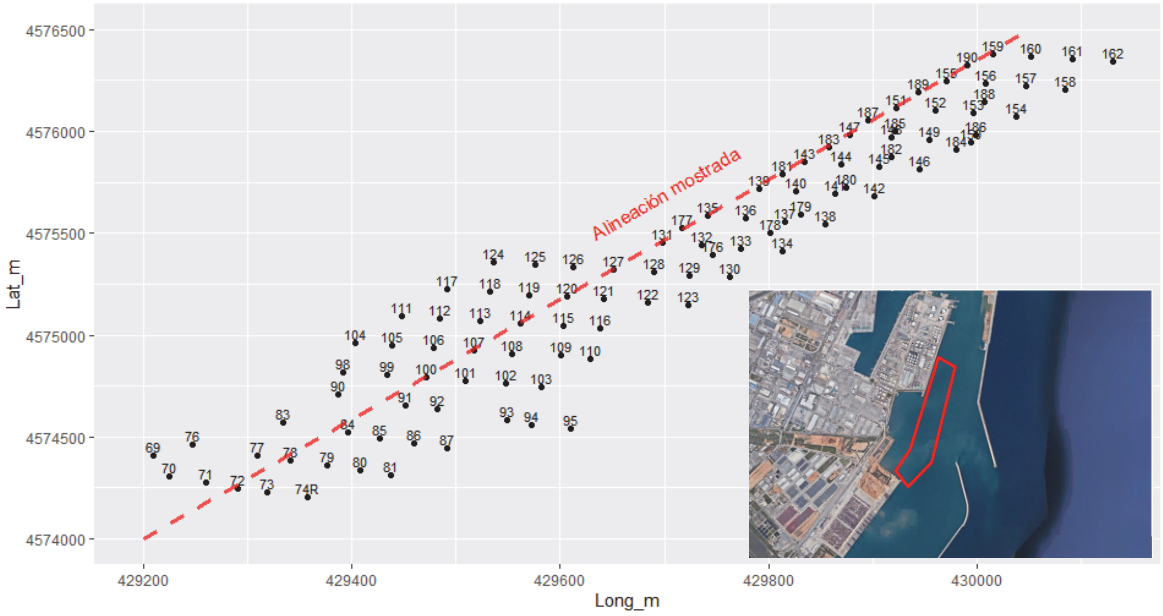


Fig. 2 – Planta de los puntos de investigación.

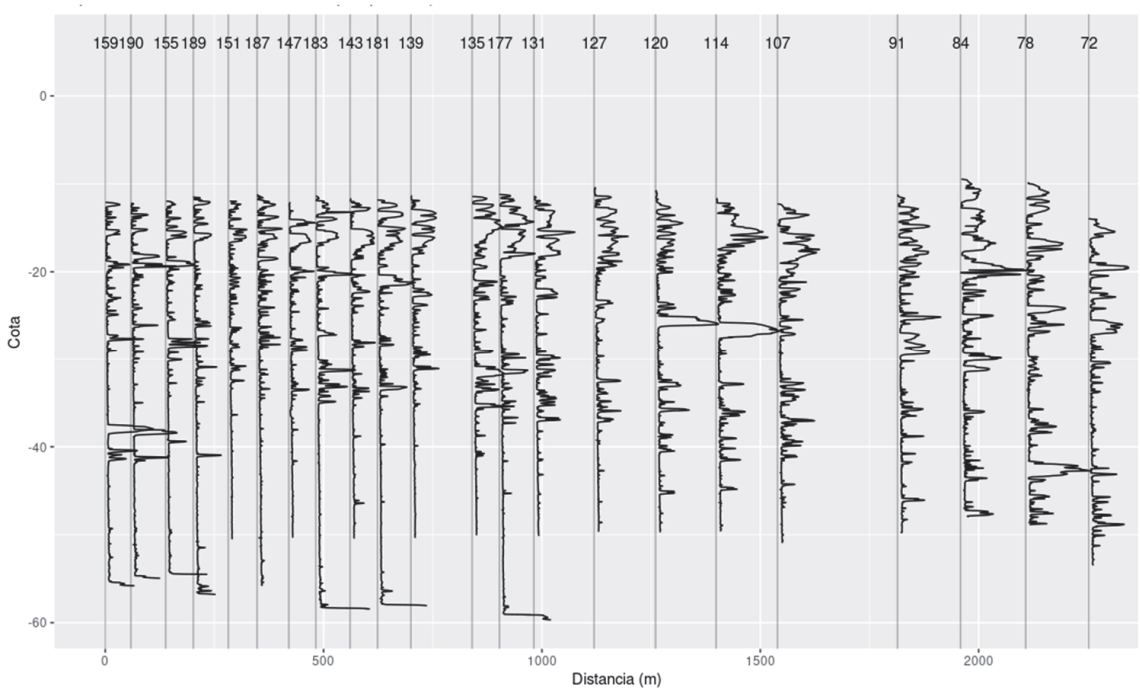


Fig. 3 – Proyección de los CPTU sobre la alineación indicada.

Para este caso de aprendizaje no supervisado se han utilizado como variables de entrada los registros normalizados de la resistencia por punta, fuste y sobrepresión de poros (Q_{tn}, Fr, B_q),

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

según Robertson (2016). Los dos datos de coordenadas (latitud y longitud), se reducen a uno sólo ("distancia") al efectuar la proyección bidimensional.

En la elaboración de los modelos se ha asumido un pre-procesado básico de datos, eliminando registros incongruentes o no representativos asociados a contingencias durante la ejecución. Se trata de un procedimiento habitual que no produce variaciones apreciables porque no elimina una cantidad significativa de registros.

5 – MODELOS PROPUESTOS Y EJEMPLOS

El proceso de caracterización que se sigue se sintetiza en dos etapas:

1. Aprendizaje no supervisado: se recoge la información y se agrupan los registros en distintas categorías homogéneas y diferenciadas.
2. Aprendizaje supervisado: Usando las variables espaciales (Coordenadas X, Y, Z) como variables predictoras y las categorías antes definidas como variables objetivo, se crea un modelo (tridimensional) que ubique espacialmente cada categoría, estableciendo las fronteras entre cada grupo de datos (como se ha aclarado, aquí se representará en dos dimensiones para facilitar la visualización)

5.1 – Aprendizaje no supervisado

La fase de aprendizaje no supervisado consiste en agrupar los datos de entrada en categorías o grupos excluyentes (unívocos), de tal forma que a cada vector m-dimensional (en el caso que nos mueve, sería un vector tridimensional de variables $\{Q_{tn}, F\mathrm{r}, B_q\}$, provenientes del registro del CPTU), se le asigne una categoría (por ejemplo, "Arenas limpias", "Mezclas arenosas-limos arenosos", "Arcillas-Arcillas limosas", etc. o, para simplificar, "A", "B", "C", etc.), con lo que se dispondría de grupos de datos categorizados.

A modo exclusivo de ejemplo, si se tuvieran ciertos resultados de límite líquido (*LL*) e índice de plasticidad (*IP*), necesarios para su representación en el ábaco de Casagrande, se puede dar una situación como la de la Figura 4. En esta situación es habitual realizar intuitivamente una agrupación de estos pares de datos (vector bidimensional {Límite líquido, Índice de Plasticidad}) en dos conjuntos como los mostrados.

Esto llevaría (a falta de mayor información) a clasificar el terreno del cual se han extraído las muestras, y sobre las cuales se han realizado ensayos de laboratorio, en dos categorías (Arcillas de alta plasticidad y Limos de alta plasticidad).

Sin embargo, el mecanismo mental por el cual se llega a dicha agrupación no parece obvio y en el proceso de describir dicho criterio se puede cuestionar si esta agrupación es óptima o se podría haber llegado a otras distintas, como las indicadas en la Figura 5, que pudieran representar de manera más eficaz, si así fuera, el número de grupos subyacentes en la información existente. Cuestionar esto lleva de manera lógica a preguntarse por qué no es tan sencillo, o unívoco, definir grupos si los datos fueran los que se muestran en la Figura 6.

Este proceso aquí simplificado se puede extender a N dimensiones. Supóngase, por ejemplo, que se dispone del N_{60}, el porcentaje de finos y el *LL*, provenientes de una muestra de un ensayo SPT. Las técnicas de aprendizaje no supervisado proponen una amplitud de criterios para establecer distintas formas de agrupar o "etiquetar" dicha información.

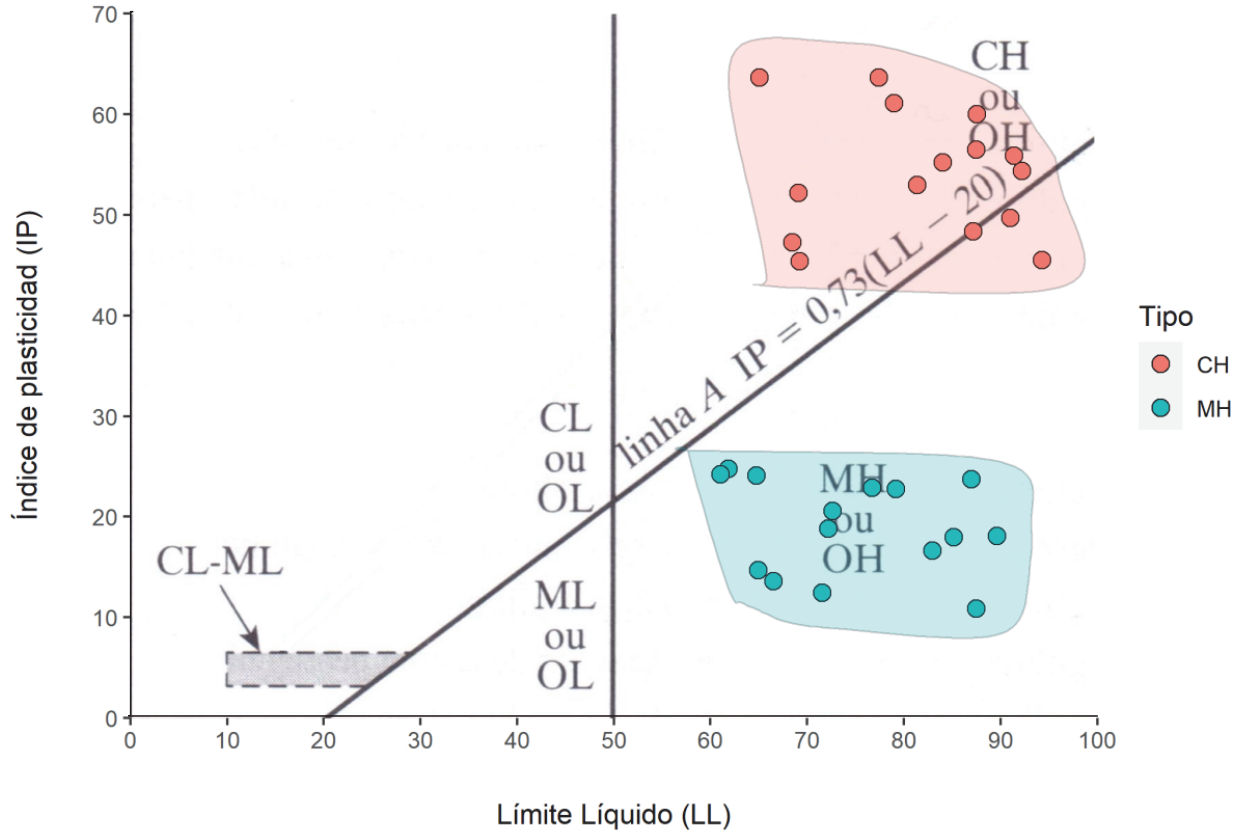


Fig. 4 – Grupos afines de valores de los límites de Atterberg en el ábaco de Casagrande

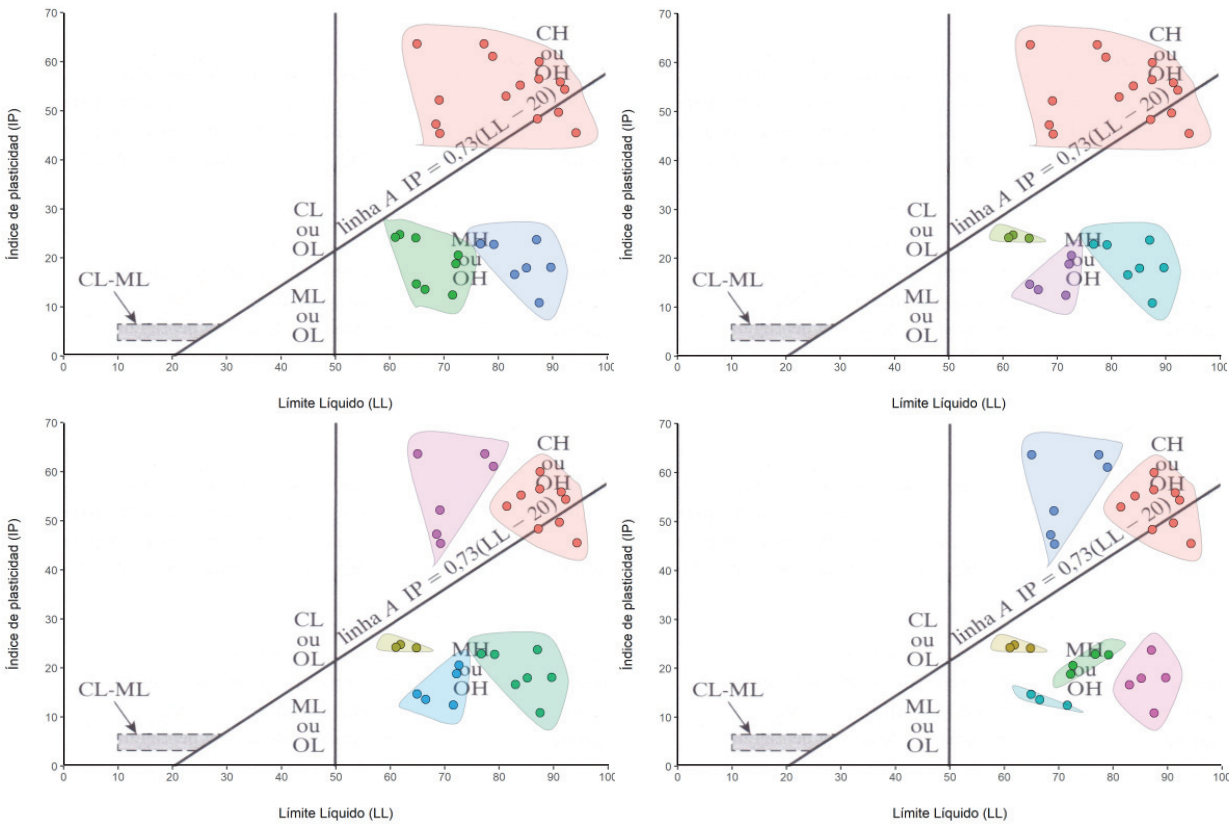


Fig. 5 – Distintas formas de agrupar los mismos pares de valores de LL e IP

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

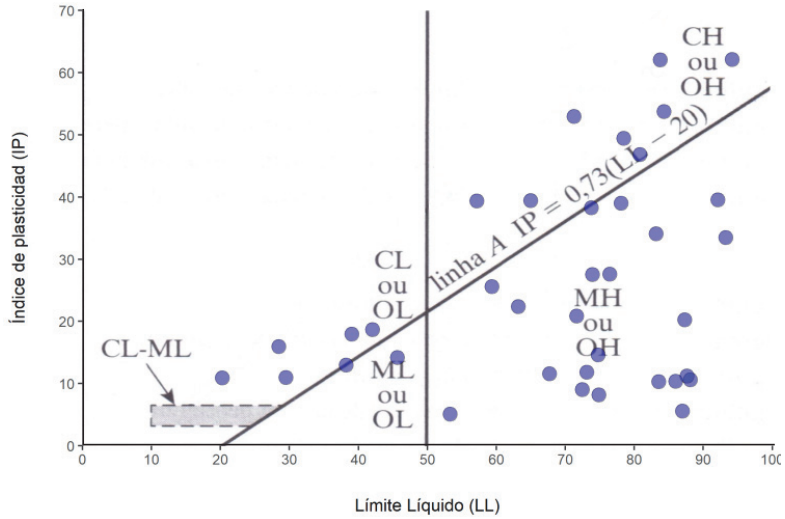


Fig. 6 – Valores de LL e IP distribuidos aleatoriamente y agrupación compleja

5.1.1 – Análisis de conglomerados (Cluster analysis) con el algoritmo k-medias (k-means)

De los muy diversos métodos o técnicas que pueden ser utilizadas para el propósito de "etiquetar" o discriminar/clasificar un conjunto de datos en diferentes "conglomerados" el probablemente más antiguo, de sencilla aplicación, comprensión y más difusión es el algoritmo de *k*-medias (MacQueen (1967), Hartigan and Wong (1979), Lloyd (1982)), usado ampliamente también en geotecnia (Wierzbicki et al (2016), etc.). Este método agrupa los datos de forma rápida y eficaz, habitualmente en muy pocas iteraciones en un número predeterminado, k, de grupos.

Sea una muestra de n observaciones y q variables. Tómense por caso, los resultados de una campaña donde se dispone de ensayos SPT, índices de plasticidad y porcentaje de finos de las muestras de cada SPT. El número de muestras se correspondería con el número n de observaciones y el número de variables sería tres (N_{60}, *IP*, %finos). El algoritmo consta ºde las siguientes 5 etapas:

1 *Seleccionar k puntos aleatoriamente como centroides de los grupos originales*: Existen diversos criterios para elegir los k centroides iniciales. A efectos prácticos se puede asumir que se toman aleatoriamente.
2 *Calcular la similitud (o distancia) de cada observación a los k centroides* anteriores y asignar cada elemento a su centroide más próximo. Comúnmente se adopta la distancia euclídea, aunque pueden adoptarse otras medidas de similitud que permitan ordenar de manera jerarquizada todos los elementos. En definitiva, se requiere un "operador" que compute el grado de "similitud" o "proximidad" entre dos puntos observaciones, y que cumpla las condiciones de distancia para poder ordenar todas ellas. La distancia euclídea se adopta aquí por ser fácilmente interpretable y visualizable, no obstante, se puede consultar bibliografía especializada que sostienen la idoneidad de otras medidas de distancia en el ámbito de la geotecnia (Wierzchon et al (2018), Hegazy et al (2002), Mlynarek et al (2005)).
3 *Recalcular el centroide* tras la nueva asignación. El valor medio de cada grupo se toma como nuevo centroide dentro de cada iteración.
4 *Definir criterio de optimización y comprobar si reasignando alguno de los elementos, mejora el criterio.*
5 Si no es posible mejorar el criterio de *optimización*, terminar el proceso.

El criterio de *optimización* utilizado en el algoritmo usado es el de minimizar la varianza intragrupal o, lo que es equivalente, la suma de las diferencias respecto al valor medio al cuadrado dentro de los grupos (Suma de Cuadrados Dentro de los Grupos):

$$SCDG = \sum_{g=1}^{k} \sum_{j=1}^{q} \sum_{i=1}^{n_g} \left(X_{ijg} - \bar{X}_{ijg}\right)^2 \quad (1)$$

siendo x_{ijg} el valor de la variable j en la observación i del grupo g, $\bar{X}_{jg}$ la media de esa variable en dicho grupo y n_g es el número de elementos dentro del grupo g. Este criterio de optimalidad es equivalente a minimizar la suma ponderada (por el número de elementos u observaciones) de las varianzas de las variables de cada grupo.

Se asumirán 3 grupos en la campaña referida, siendo las variables Q_{tn}, B_q y Fr(%). La Figura 7 muestra los 3 grupos resultantes sobre los ábacos de Robertson (2016), junto los 3 grupos en el espacio tridimensional (no físico) Q_{tn}, B_q y Fr(%). La Figura 8 muestra la visualización de los valores en el modelo tridimensional de la campaña

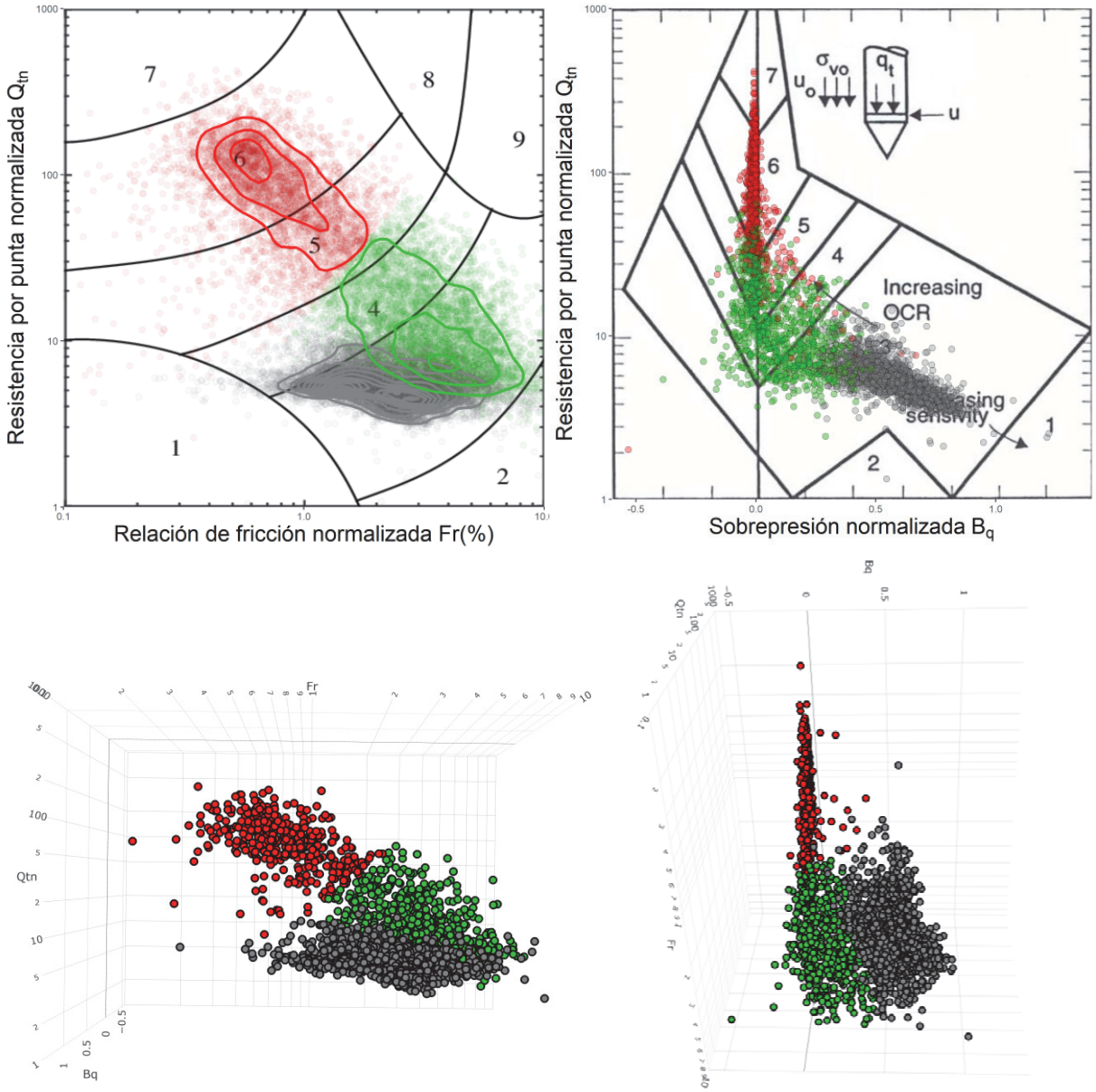


Fig. 7 – (Arr.) Agrupación de datos (en 3 grupos) obtenidos mediante el algoritmo de *k*-medias sobre los ábacos de Robertson (2016). (Abaj.) Representación en el espacio tridimensional Q_{tn}, Fr % y B_q agrupados por clases.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

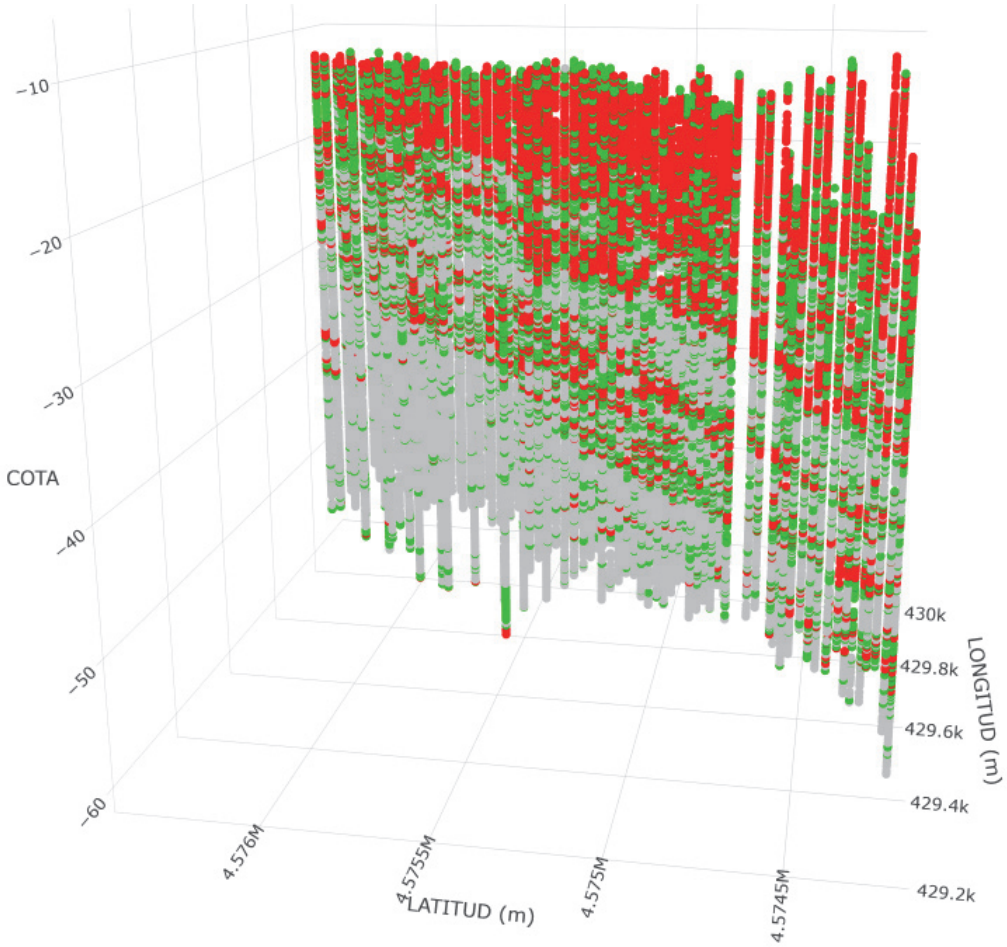


Fig. 8 – Vista 3D de los registros de los CPTUs agrupados.

La normalización de las variables es habitual en el proceso de creación de conglomerados, pues las distancias quedan desligadas de las unidades físicas de las variables en bruto. La normalización univariante consiste en realizar la transformación lineal de los valores de cada variable restándoles su valor medio (centrado) y dividiéndoles por su desviación típica (estandarizado), de tal forma que todos los valores puedan ser representados por el número de desviaciones típicas que se alejan respecto a su valor medio. Aunque en general es recomendable, la conveniencia o no de esta transformación debe ser estudiada en función del proceso y datos que se busque analizar.

El algoritmo de *k*-medias es simple de aplicar y visualizar, muy rápido, escalable e intuitivamente comprensible. Valores extremos o atípicos ("outliers") muy acentuados, aunque sean escasos, tienden a generar grupos independientes, lo que permite por un lado, su detección, y por otro, si procede, la evaluación de su efecto. En el caso de los registros de los CPTUs, tiende a formar grupos con las ternas de valores eliminados en el pre-procesado.

Por otro lado, la propiedad principal del criterio de distancias asumido es que resultan grupos aproximadamente esféricos y excluyentes. Cuando los datos no formes grupos pseudo-esféricos, o tengan acentuados diferentes tamaños o densidades, este algoritmo podría dar lugar a agrupaciones erróneas. En este sentido, el conocimiento previo del proyectista sobre el resultado esperable debe hacerse valer. Los resultados deben contrastarse visualmente para su validez y, si procede, someterse a un algoritmo distinto.

Dado que el algoritmo es sensible a la posición de los centroides iniciales, puede corregirse de varias formas; una de ellas consiste en repetir varias veces el proceso de inicialización aleatoria para adoptar los grupos resultantes óptimos de entre todas las inicializaciones.

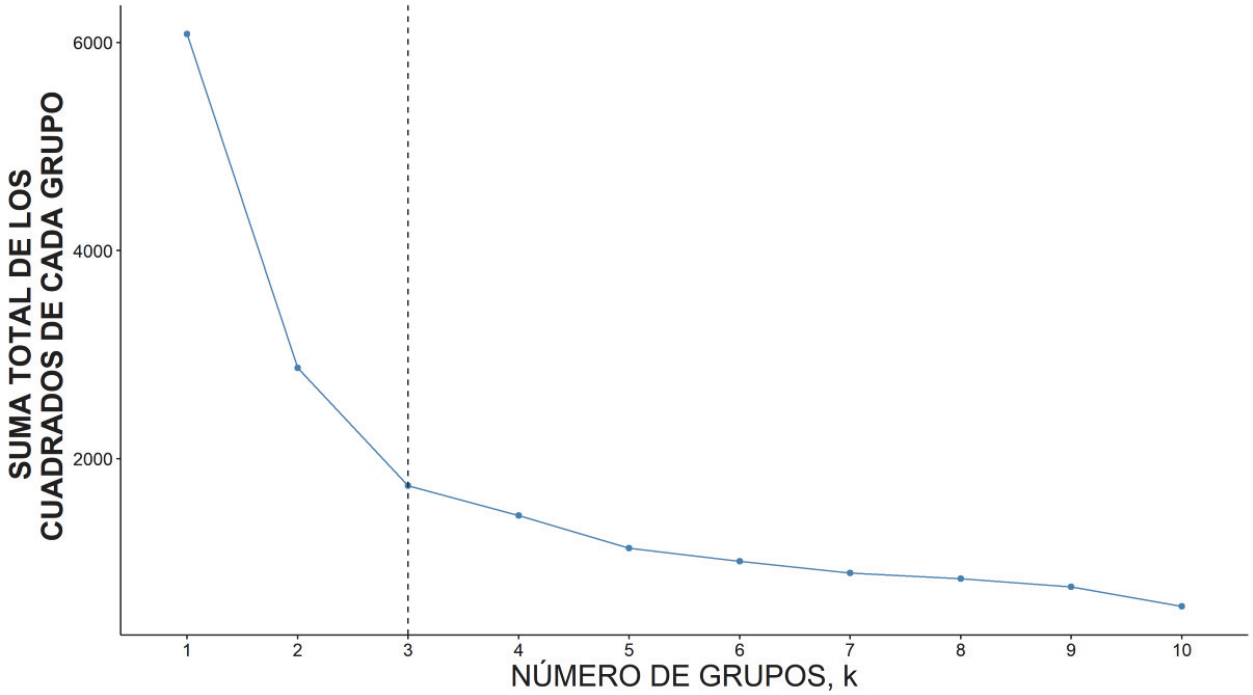


Fig. 9 – Balance entre el número de grupos y el criterio de optimización para los datos de la alineación.

Un inconveniente del algoritmo de *k*-medias (y otros) es que precisa una postura a priori sobre el número de grupos a definir. Esta postura no se puede definir con un criterio de optimización (simple), pues la varianza intragrupal disminuirá conforme aumente el número de grupos establecido. En rigor, la situación óptima sería un número de grupos igual al número de elementos disponible, en el que la varianza dentro de cada grupo sería nula y, por lo tanto, el sumatorio resultaría en SCDG=0. Típicamente, el sumatorio de varianzas varía con el número de grupos según una curva similar a la de la Figura 9. Se trata de establecer un equilibrio entre un número de grupos no excesivo (evitando definir grupos inexistentes) y una varianza intragrupal pequeña (que haga que cada grupo contenga información concreta). De forma simplista, este equilibrio se puede establecer en el "codo" de la mencionada curva. Een, también, otros procesos más elaborados para la determinación del número de grupos. Los resultados que se exponen aquí (3 grupos), se han obtenido con esta "regla del codo".

5.1.2 – Análisis de conglomerados (Cluster analysis) con el algoritmo PAM (Partitioning around medoids).

El algoritmo de *k*-medoides (traducción anglicista del término "medoid") es similar en muchos sentidos al de *k*-medias. El "centro" (o valor representativo) de cada grupo está restringido a ser alguna de las observaciones o elementos del grupo. Para cada grupo (o conglomerado) ha de encontrarse el elemento que hace que la suma de las distancias de todos los elementos del grupo a dicho elemento "central" sea mínima y, una vez definido este "centro", ha de reasignarse cada elemento a su "centro" más cercano, procediendo de forma iterativa hasta que no se produzcan más reasignaciones.

A priori supone un coste computacional mucho mayor que con el algoritmo *k*-medias, pues implica calcular muchas más distancias en el paso de cálculo del centro de cada grupo. Tiene, no obstante, dos ventajas fundamentales: en primer lugar, el algoritmo de *k*-medoides se resuelve con el algoritmo de Partición Alrededor de Medoides (Partitioning around medoids, PAM) que reduce significativamente el coste computacional. Este consiste en iniciar con tres (o "*k*") puntos

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

"centrales" (medoides), asignar a cada punto su medoide más próximo y calcular la matriz de distancias de cada grupo. Si cualquiera de los elementos del grupo k (asumido como centroide) disminuye la función de coste, se cambia por el centroide anterior, y se vuelven a asignar elementos a cada centroide, procediendo de forma iterativa hasta que ningún "medoide" se vea modificado. En segundo lugar, el algoritmo de k-medoides puede ejecutarse también con otro, denominado CLARA, para agrupar registros de gran tamaño (Clustering Large Applications). El algoritmo opera con una sub-muestra del conjunto inicial ejecutando el algoritmo PAM sobre ella. Posteriormente, asignar a cada centroide el resto de elementos de la muestra original. Se realiza este procedimiento en diversas ocasiones para adoptar, finalmente, la agrupación que optimice la función de coste (en este caso la suma de distancias de todos los elementos a su medoide).

En el caso de la citada campaña, con el pre-procesado realizado, los resultados son similares con uno y otro algoritmo (Figura 10).

El algoritmo es robusto y resulta una buena alternativa al *k*-medias menos sensible a valores atípicos.

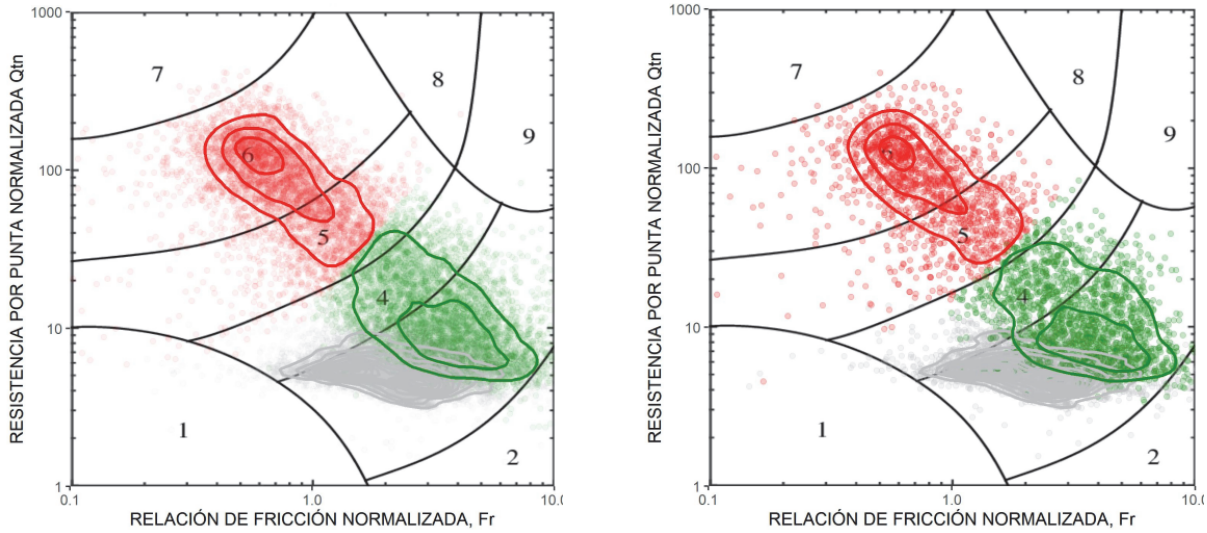


Fig. 10 – Resultados de los algoritmos: (Izq.) *k*-medias; (der.) k-medoides (CLARA).

5.1.3 – Agrupamiento de mezclas de distribuciones. Algoritmo de Esperanza-Maximización (EM- Expectation Maximization algorithm).

Si en lugar de asumir que las observaciones son elementos aislados, pertenecientes a grupos excluyentes y cuya información es el grupo al que pertenecen y la terna de valores que los definen, se consideran pertenecientes a distribuciones estadísticas de *n*-dimensiones, se puede plantear como un problema de mezcla de distribuciones. Su análisis parte de la idea de que si queremos dividir una población en *k* grupos más homogéneos, la distribución conjunta vendrá dada por la función de mezcla de densidades de la forma:

$$f(x) = \sum_{i=1}^{k} \pi_i \cdot f_i(x) \qquad (2)$$

Siendo $f_i(x)$ la función de densidad del grupo "i" y π_i la proporción de elementos en ese grupo "i", de tal manera que el sumatorio de estas proporciones sea la unidad. Esta idea viene justificada por el hecho de que una observación se puede plantear en dos pasos: primero seleccionar el grupo del que se quiere extraer el elemento (π_i se puede entender como la probabilidad de elegir el grupo

"i") y, segundo, elegir, de forma aleatoria, un elemento de ese grupo. Típicamente se tratan de modelos gaussianos multivariantes,

$$N(x;\ \mu, \sum)\ =\ \frac{1}{(2\pi)^{d/2}}\ |\sum|^{-1/2} exp\left\{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right\} \tag{3}$$

Y, en el caso particular aquí expuesto, x sería un vector tridimensional de variables:

$$x_i = \{Qtn_i, Fr_i, Bq_i\} \tag{4}$$

El concepto de verosimilitud resulta de invertir el objeto de la función de densidad, asumiendo conocida la muestra y la forma de la distribución, pero desconocidos los parámetros característicos (μ, Σ) que la definen. A partir de la función de densidad conjunta de una muestra específica (producto de las probabilidades individuales), se puede obtener la función de verosimilitud de la que los parámetros característicos (μ, Σ) son las variables, y los elementos que la componen, fijos. Al maximizar esta función respecto a los parámetros que la definen, conociendo los elementos que la componen, se obtienen los parámetros de máxima verosimilitud: los más verosímiles para la población que contienen. Luego, se pueden estimar los parámetros máximo-verosímiles a partir de la distribución normal multivariante.

El denominado algoritmo EM (Hastie et al, 2008), o de Esperanza-Maximización (Expectation – Maximization) es una herramienta muy habitual para el cálculo de la máxima verosimilitud de mezclas de distribuciones mediante un algoritmo iterativo. Se puede aplicar a cualquier tipo de distribución, pero en el presente caso se aplica y se expone asumiendo distribuciones normales. Típicamente, se ejecuta en dos etapas, denominadas E (Esperanza) y M (Maximización). En la primera (E) se asigna a cada elemento la probabilidad de pertenecer a cada una de las distribuciones iniciales de la función de densidad conjunta con la que se inicializa el proceso. De esta forma se asigna a cada elemento, una distribución. Posteriormente, con la asignación obtenida, se actualizan los parámetros que definen las distribuciones, calculando π_i, μ_i y Σ_i. Se demuestra que con este proceso la función de soporte, el logaritmo de la función de verosimilitud, aumenta con cada iteración y garantiza la convergencia.

El algoritmo EM para mezclas gaussianas y el agrupamiento de *k*-medias están íntimamente relacionados, pudiéndose entender éste último como un caso particular de mezcla de distribuciones. También resulta de interés para estimar valores ausentes y porque asigna una probabilidad a cada elemento por hacerlos pertenecer a una distribución. Permite, además, agrupamientos más flexibles que el *k*-medias, el cual tiende a agrupar en esferas *n*-dimensionales. Los resultados de este agrupamiento con el algoritmo EM se muestran en la Figura 11.

La convergencia no garantiza alcanzar un óptimo global, por lo que, al igual que en el algoritmo de *k*-medias, requiere varias inicializaciones y usar la función de soporte, mediante algún criterio de selección de modelos, para obtener el mejor de ellos. No obstante, y dependiendo del tipo de terreno y obra a desarrollar, puede ser de interés profundizar en el número de grupos o "estratos" a considerar, típicamente mediante agrupamientos jerárquicos y algún evaluador o combinación de ellos.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

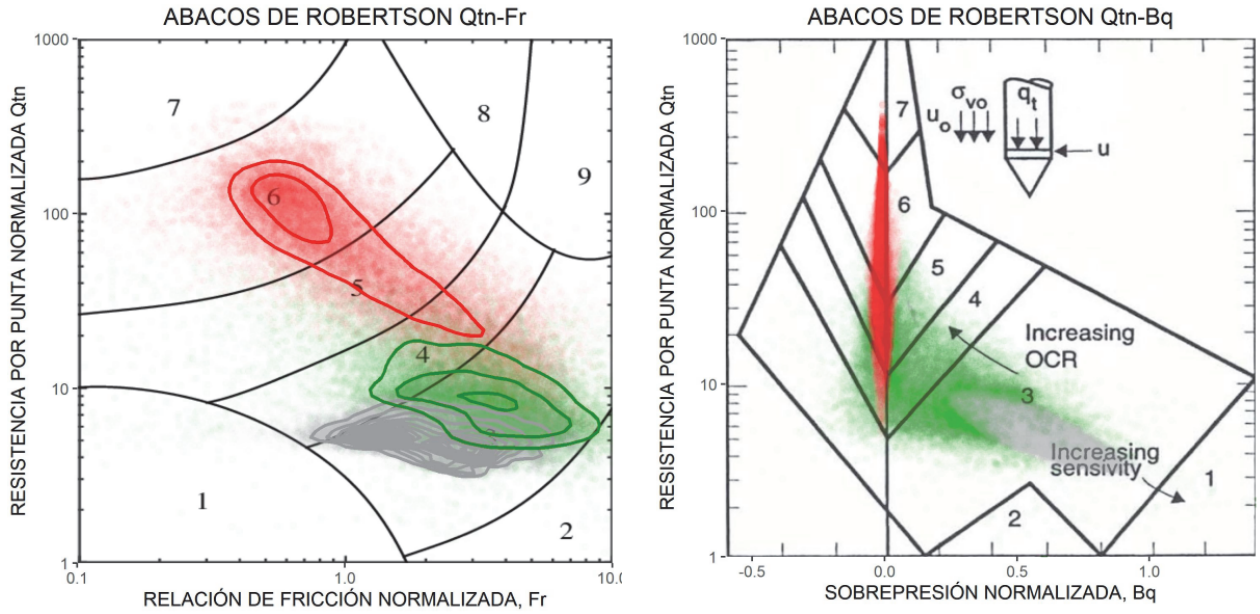


Fig 11 – Grupos de datos obtenidos mediante el algoritmo de EM, asumiendo 3 grupos representados sobre los ábacos de Robertson (2016).

5.2 – Aprendizaje supervisado

Ya agrupados los datos con cualquiera de los criterios anteriormente descritos, se dispone de un conjunto de datos categorizados. Con esta información se debe elaborar alguna estrategia que permita asignar una de estas categorías a cualquier conjunto de variables. En este caso se dispone de una serie de puntos distribuidos en el espacio tridimensional (o bidimensional si se proyecta en 2D). Las variables serán, por tanto, *X*, *Y* y *Z*; y a cada punto disponible le corresponde una de las categorías anteriormente definidas. En el perfil de la Figura 3, usando el algoritmo EM, cada punto ensayado (tres valores por centímetro) tendría asignado un grupo, según se ve en la Figura 12. Esta es la información de la que se parte para la etapa de aprendizaje supervisado.

La elaboración de un perfil o la compleción del espacio tridimensional (o bidimensional en el caso de un perfil) consistirá en asignar a cualquier punto del espacio sin clasificar (todo el fondo gris) una categoría siguiendo algún criterio objetivo. Aunque se han desarrollado diversos criterios, en este artículo se expondrán solo cuatro: *k*-vecinos más próximos, máquinas de vector soporte, bosques aleatorios y redes neuronales.

En todos los casos la primera etapa consiste en dividir el espacio en una malla de puntos lo tupida o gruesa que se considere; a continuación, se asigna una clase (o grupo) a cada uno de ellos (Figura 13).

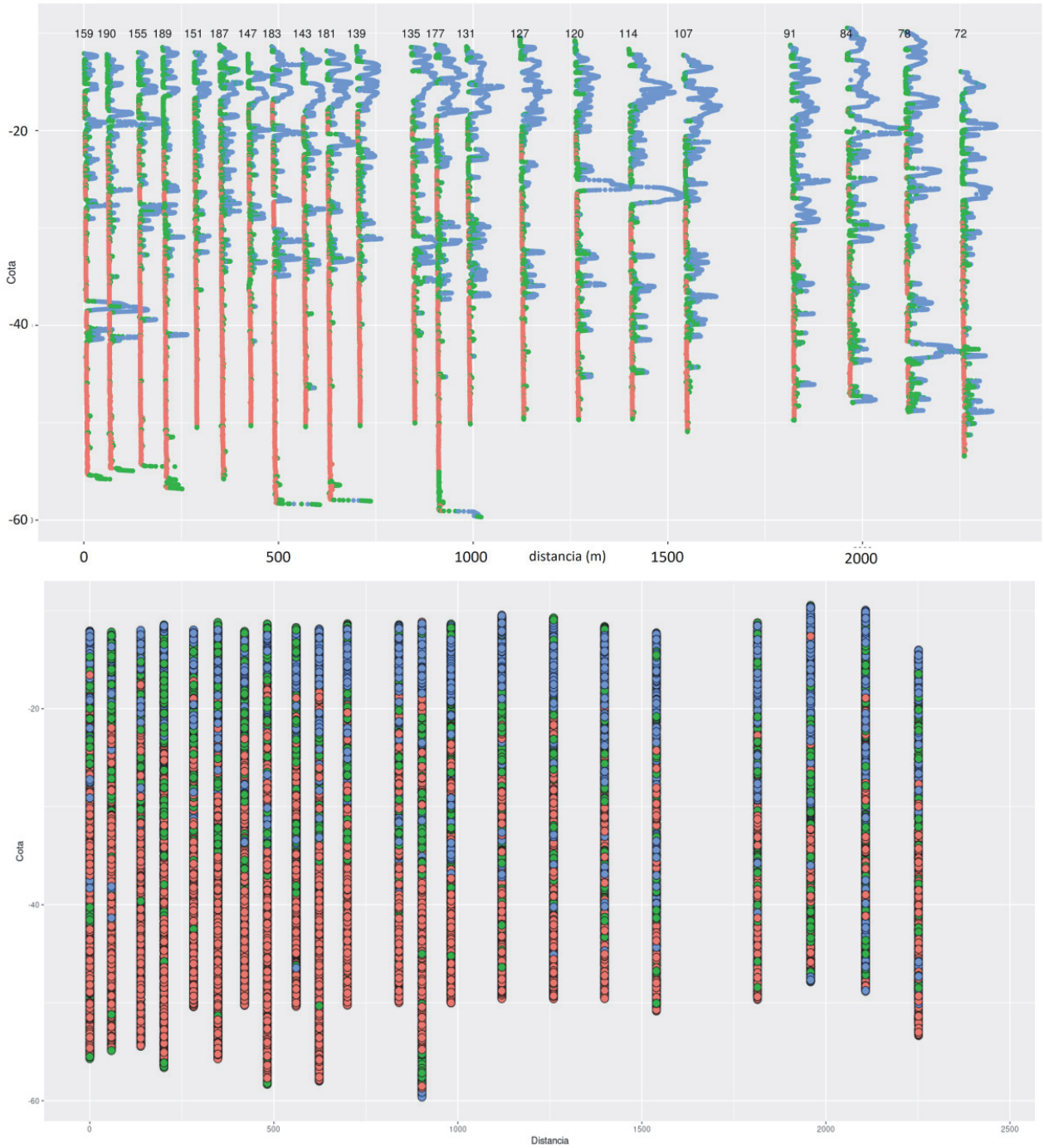


Fig. 12 – Resultados de grupos encontrados con el algoritmo E-M: (arr.) resistencia por punta; (ab.) representación de cada clase en la proyección bidimensional.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

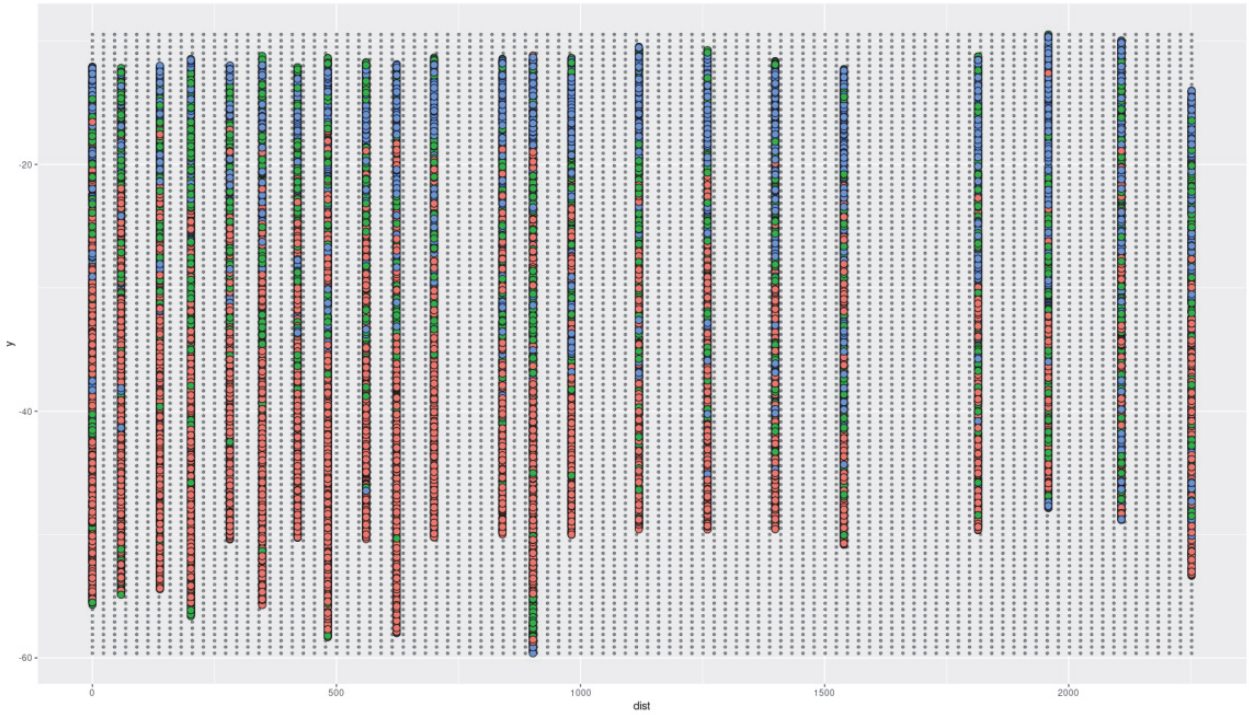


Fig. 13 – Malla de puntos; en este caso, de 100x100 puntos, pendientes de asignar un grupo.

5.2.1 – Validación cruzada

El proceso denominado "validación cruzada" ("cross-validation", Hastie et al, 2008) es aquel en virtud del cual se autoconfirma, o evalúa, la validez de la solución adoptada con los datos de prueba. No se detalla aquí, pero consiste en usar uno o varios subgrupos, de forma secuencial, de la muestra categorizada y usarlos para validar la conformidad del modelo que se desarrolle con el resto del registro. La validación cruzada, que puede efectuarse de diversos modos, repercute enormemente tanto en el modelo resultante como en el coste computacional.

Más allá de las distintas formas de validación cruzada que se suelen aplicar, la particularidad geotécnica de extraer registros completos de puntos de investigación específicos ha llevado a evaluar este proceso comparando la validación, bien mediante un porcentaje de puntos seleccionados de forma aleatoria de entre el grupo total de registros o, por el contrario, mediante la extracción de los registros completos correspondientes a uno o varios de los puntos de investigación. Ambas formas de aproximar el problema llevan soluciones claramente distintas y, quizá, suponen un ámbito de estudio en sí mismo. Esta forma de validación está íntimamente ligada con el nivel de simplificación que se está dispuesto (o intencionadamente se busque) a adoptar.

Resulta difícil de prefijar el grado de simplificación que puede admitirse a un modelo geológico o geotécnico para su uso en ingeniería civil; depende específicamente de la estructura o diseño del que se trate, así como de las peculiaridades del terreno. Por eso, modelos definidos por su buena capacidad predictiva resulta en perfiles geotécnicos poco prácticos. Es responsabilidad del diseñador, en función del objetivo con el que se busque desarrollar el modelo y las particularidades propias del diseño, el grado de simplificación del modelo que se busque y se deberá reconocer a partir de distintos evaluadores o de la matriz de confusión (Hastie et al, 2008).

5.2.2 – Evaluadores

Existen muy diversos evaluadores o métricas para evaluar la bondad o adecuación del modelo (Khun y Johnson, 2016). Suponen también un campo de estudio en sí mismo, por lo que no se

profundiza. En modelos complejos y de gran alcance es prescriptivo evaluar de forma conjunta varios evaluadores de forma rigurosa. En los modelos aquí descritos se ha usado, por su valor intuitivo, la kappa de Cohen para clases múltiples como métrica para evaluar los modelos. Su utilidad radica en que representa un valor equivalente al error de clasificación ("accuracy") pero considerando los valores que el modelo acertaría por azar. Supóngase un terreno en que el 95% del perfil son arenas y el 5% son arcillas. Un modelo que predijera que todo el terreno son arenas tendría una precisión del 95% (acertaría un 95% de las veces), pero sería un mal modelo porque no detectaría jamás la capa de arcillas. Pueden existir capas que, estando poco representadas en el terreno, gobiernen el comportamiento de una cimentación, como una lámina de poco espesor de arcillas muy sensitivas. Por ello se usa el coeficiente de la expresión siguiente:

$$Kappa = \frac{Obs - Prev}{1 - Prev} \tag{5}$$

donde "*Obs*" es el error observado (la probabilidad de que el modelo acierte que la clase predicha es la correcta), y "*Prev*" es el valor esperado basado en los marginales totales de la matriz de confusión (que podría representar la probabilidad de acertar con las clases estimadas usando valores al azar). Si bien no hay una regla evidente, los valores de *Kappa* entre 0.3 o 0.5, se pueden considerar ajustes razonables. En cualquier caso, el *Kappa* medio es sólo uno de los posibles evaluadores a utilizar.

5.2.3 – Regularización

La regularización es un proceso importante en los modelos supervisados (Hastie et al, 2008). Su objetivo es penalizar de alguna manera la estimación realizada para evitar sobreajustarse a la información disponible.

Existen muy diversas formas de regularización. En modelos como el de *k*-vecinos más próximos o los bosques aleatorios se tiende a solventar el sobreajuste con la validación cruzada. Sin embargo, modelos como los de redes neuronales o máquinas de vector soporte, incluyen términos de penalización en las funciones de coste.

5.2.4 – Método de los k-vecinos más próximos (k-nearest neighbours)

Este criterio, por su simplicidad, rapidez y porque, intuitivamente, podría equipararse al proceso básico que el ojo humano repite en la elaboración de un perfil, se considera muy apropiado, si no como modelo definitivo, sí como una primera aproximación visual al perfil aproximado. Esquemáticamente el proceso de asignación es como sigue:

1- Una vez elegida la medida de distancia a usar (en este caso, la euclídea), se calcula la distancia entre cada punto sin categorizar y sus *k* puntos más próximos de clase conocida.
2- De entre los *k* puntos obtenidos, se estima la proporción de puntos de cada clase.
3- Se asigna al punto sin categorizar la clase mayoritaria entre los *k* puntos más próximos.

A priori se desconoce el valor de *k* que mejor define el modelo. Se opera, por tanto, con un rango amplio para el valor de *k*, para así poder, posteriormente, juzgar mediante validación cruzada cuál es la mejor aproximación. En la campaña descrita se operó con *k* entre 1 y 400.

Los resultados se pueden analizar en función del número de subgrupos para la validación cruzada y el tipo de esta validación (aleatoria, adoptando un subgrupo aleatorio del conjunto de lecturas, o por punto de investigación, extrayendo el registro completo de uno o varios de los puntos investigados como subgrupo de validación). Según muestra la Figura 14, si se realiza una validación cruzada tomando valores de forma aleatoria, el valor de *kappa* tiende a disminuir de forma paulatina desde un valor máximo en *k*=1. Es decir, tomado un punto al azar, lo más probable

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

es que tenga la misma categoría que el punto que se encuentre más próximo a él. Sin embargo este tipo de modelos conduce a representaciones poco prácticas y visualmente rechazables, parecidas a las obtenidas con el modelo de bosques aleatorios pese a que, objetivamente, pudieran tener mejor capacidad predictiva.

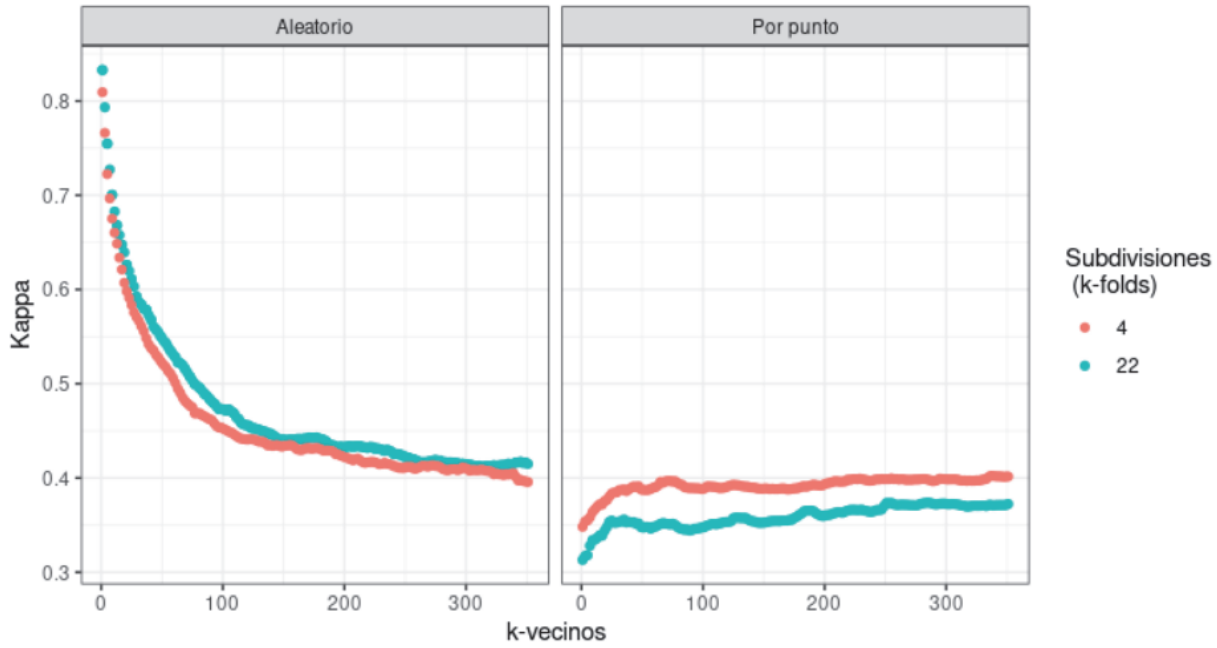


Fig. 14 – Variación de la kappa de Cohen en función del número k y tipo de validación.

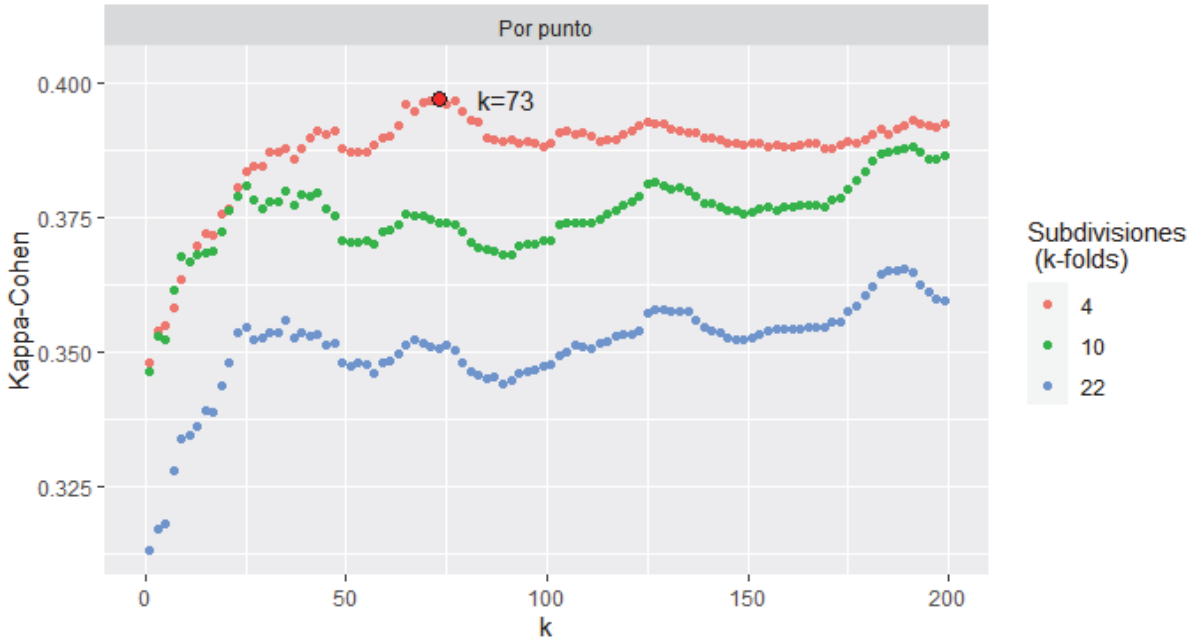


Fig. 15 – Variación de la kappa de Cohen en función de k para tres valores de subdivisiones, empleando la validación por punto de investigación.

Sin embargo, en el caso de validación por punto de investigación (Figura 15) el valor de kappa tiende a ascender con el número *k* de vecinos hasta un valor en torno a *k*=60 y, en adelante, mantiene un nivel aproximadamente estable. De alguna manera, adoptar una validación cruzada extrayendo los registros por punto de investigación favorece la representación de la

subhorizontalidad natural del terreno ganando generalidad, aunque se pierda en capacidad de predicción.

Si se aumenta *k*, se gana en generalidad y se pierde en capacidad descriptiva. (Figura 16).

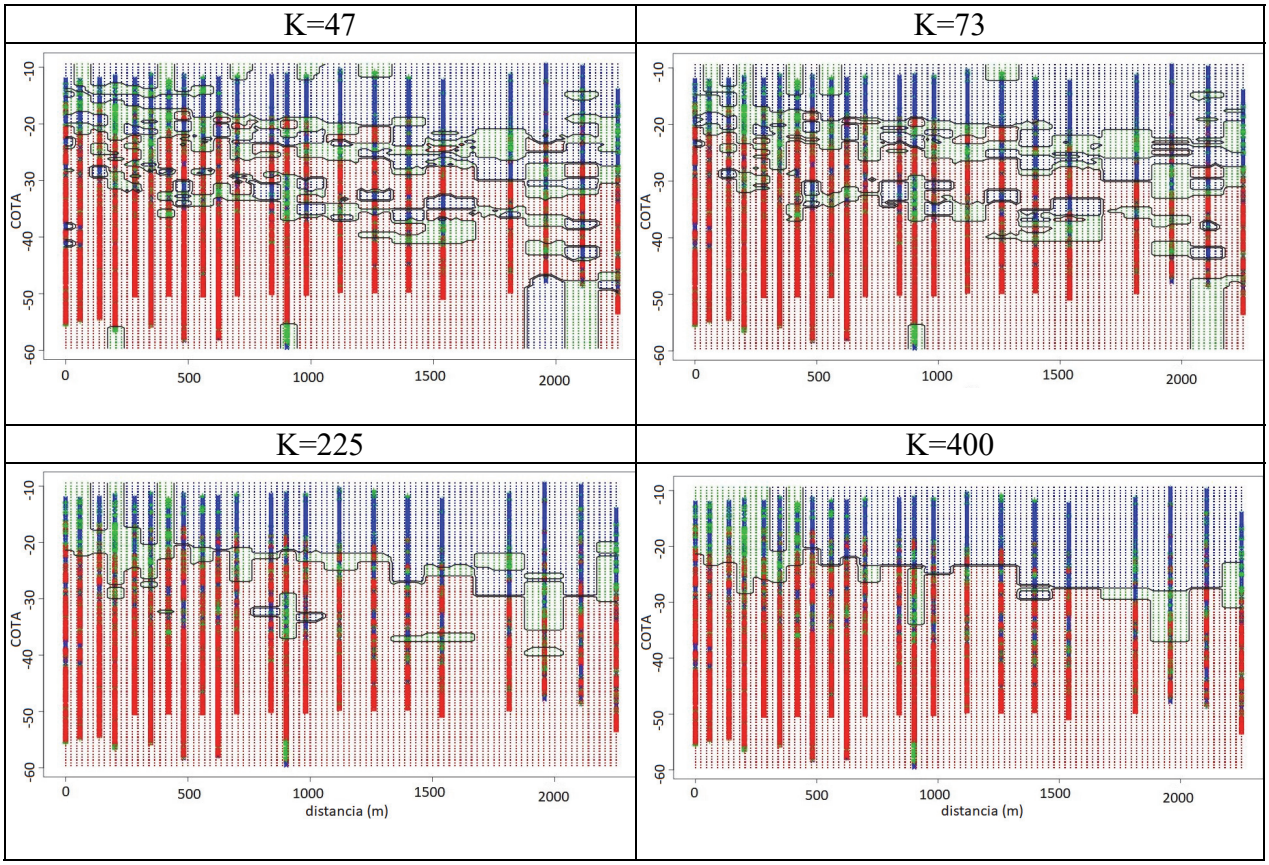


Fig. 16 – Modelo de terreno obtenido con el algoritmo de *k*-vecinos más próximos; con diferentes valores de *k* y validación cruzada por punto de investigación con 4 subdivisiones.

Estos ejemplos tienen el interés particular de aproximarse a algunas de las distintas representaciones que, históricamente, se han realizado del perfil de la zona de estudio. Antiguamente, se tendía a modelos más generales (*k* alto) y, recientemente, los modelos ganan en complejidad (*k* bajo). En cualquier caso, el valor de *kappa* no varía sustancialmente a partir del primer máximo relativo en el caso de validación cruzada por punto de investigación (Figura 15).

Se ha considerado, de forma similar a como se selecciona el número de grupos, que un valor razonable de *k* es aquel para el cual la *kappa* de Cohen no varía sustancialmente en una validación por punto ($k \approx 73$).

5.2.5 – Método de los bosques aleatorios (random forest)

El algoritmo de bosques aleatorios es, tal vez, uno de los más utilizados como modelo predictivo. Tiende a dar modelos sobreajustados y, pese a que pudieran no considerarse del todo apropiados para obtener los modelos de terreno que se buscan, se expone también aquí por las múltiples posibilidades de aplicación en otros muchos ámbitos de la geotecnia y por sus aptitudes como modelo predictivo. Se puede entender como una extensión del "bagging" para árboles de decisión para clasificación (o regresión).

La técnica de árboles de decisión (o clasificación) es una estrategia intuitiva y no estadística para clasificación. La idea de los árboles de clasificación es partir de un conjunto de datos y buscar

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

la forma de dividir este conjunto, de manera recurrente, en subgrupos, sucesivamente más homogéneos. Al aplicar este criterio de forma sucesiva se puede alcanzar el grado de homogeneidad deseado en cada subgrupo.

A modo de ejemplo, por su claridad, tómese, por caso, la base de datos de Boulanger y Idriss (2014) sobre sismos en los que se ha producido o no licuefacción. Se trata de una base de datos de 210 eventos que incluyen variables como la magnitud y aceleración del sismo, el porcentaje de finos o el SPT. A este grupo de valores se le añade una columna reflejando si se ha producido o no licuefacción.

Tabla 1 – Base de datos (incompleta) de sismos para estudio de potencial de licuefacción (Cetin et al., 2018)

AÑO	LUGAR	LICUEFAC	ACELERACION	*M*	% FINOS	$N_{1,60,CS}$
1944	Tohnankai	Yes	0.2	8.1	72.5	5.42
1944	Tohnankai	Yes	0.2	8.1	9.67	9.89
1944	Tohnankai	Yes	0.2	8.1	19.3	5.46
1948	Fukui	Yes	0.4	7	0	7.03
1948	Fukui	Yes	0.35	7	3.29	21.2
1964	Niigata	Yes	0.09	7.6	5	8.4
1964	Niigata	Yes	0.16	7.6	2	11.4
1964	Niigata	Yes	0.16	7.6	2	11.6
1964	Niigata	Yes	0.16	7.6	2	11.9
1964	Niigata	Yes	0.16	7.6	0	7.1
# … con 200 filas más….						

Aplicando la estrategia de un árbol de decisión a los tres parámetros principales (el porcentaje de finos se considera recogido en el valor de $N_{1,60,CS}$) para determinar la susceptibilidad a la licuefacción, y utilizando un proceso interno de validación cruzada, resulta el árbol de decisión de la Figura 17. El objetivo, como se ha dicho, es ir subdividiendo cada rama en sub-ramas más homogéneas, esto es, que contengan una mayor proporción de una clase en cada nodo o que, de alguna manera, aporten mayor información clasificatoria.

Se puede observar que, con este modelo, no se esperan casos de licuefacción para valores de $N_{1,60,CS}$ mayores de 26 y que para valores entre 23 y 26, apenas hay un 29% de probabilidades de licuefacción. Por el contrario, para valores $N_{1,60,CS}$ menores de 23 y sismos de magnitud superior a 6.6 las probabilidades de licuefacción son altas. Se puede prolongar la división hasta alcanzar la homogeneidad o "pureza" que se considere oportuna, o el número de elementos mínimo en cada rama. Con este tipo de estructuras se puede establecer un modelo predictivo para cualquier terna de valores de magnitud, aceleración y $N_{1,60,CS}$ siguiendo la estructura hasta su base.

Existen muy numerosas alternativas para establecer el criterio de división de cada rama. Una muy habitual, y utilizada en este caso, es el índice GINI. Para el caso simple de un problema con dos clases (Licuefacta – No licuefacta), se expresaría como:

$$GINI = p_1 \cdot (1 - p_1) + p_2 \cdot (1 - p_2) \tag{6}$$

donde p_1 y p_2 representan las probabilidades o frecuencias de la clase 1 y la clase 2 respectivamente. Se puede ver que, si se consigue que un nodo divida en dos subgrupos en los que uno de ellos sólo contiene elementos de una clase (p_1=1 y p_2=0), este valor se minimiza. Es decir, un valor de 0 indica una perfecta homogeneidad y un valor de 0.5 implica una completa falta de

homogeneidad (todas las clases están equilibradas en esa rama y no contiene información). Este formula es generalizable a n clases:

$$GINI = \sum_{k=1}^{K} p_k \cdot (1 - p_k) \tag{7}$$

Existen diversas opciones para reducir la varianza o incertidumbre de este tipo de modelos. El "bagging" es un acrónimo de "bootstrap aggregation", de difícil traducción al castellano. El término "Bootstrap" fue originalmente concebido para incidir en la autosuficiencia del método y hace referencia al dicho en inglés "*pull yourself up by your bootstraps*".

La idea fundamental es que, en caso de modelos complejos, promediando los resultados se obtienen modelos más suavizados que dan un mejor equilibrio entre el sesgo del modelo y su varianza. Es decir, se trata de realizar diversos modelos con remuestreos con sustitución de la población de que se disponga para, posteriormente, ponderarlos o promediarlos. Es una aproximación que no varía sustancialmente el sesgo del modelo, pero que tiende a reducir claramente la varianza.

La técnica de bosques aleatorios es una extensión natural de esta técnica de estimaciones autosuficientes aplicada a árboles de decisión. Se selecciona, en primera instancia, el número de árboles a construir y, para cada uno de esos árboles (generados por "bootstrapping") se construye el modelo de clasificación. Cuando el número de variables es elevado (y no es condición "sine qua non"), pueden existir relaciones subyacentes entre ellas y pudieran reflejarse en la similitud entre árboles de distintos remuestreos. En el caso de los bosques aleatorios, en cada división de una rama también se remuestrean las variables usadas. Es decir, sólo un subconjunto de las variables descriptoras se utiliza en cada división. En el caso anterior, por ejemplo, la primera subdivisión se podría hacer usando sólo las variables "aceleración" y "$N_{1,60,CS}$" y, para la siguiente subdivisión, hacer uso únicamente de la aceleración y la magnitud. Se genera, finalmente, con este procedimiento, un número elevado de árboles y se ponderan o promedian.

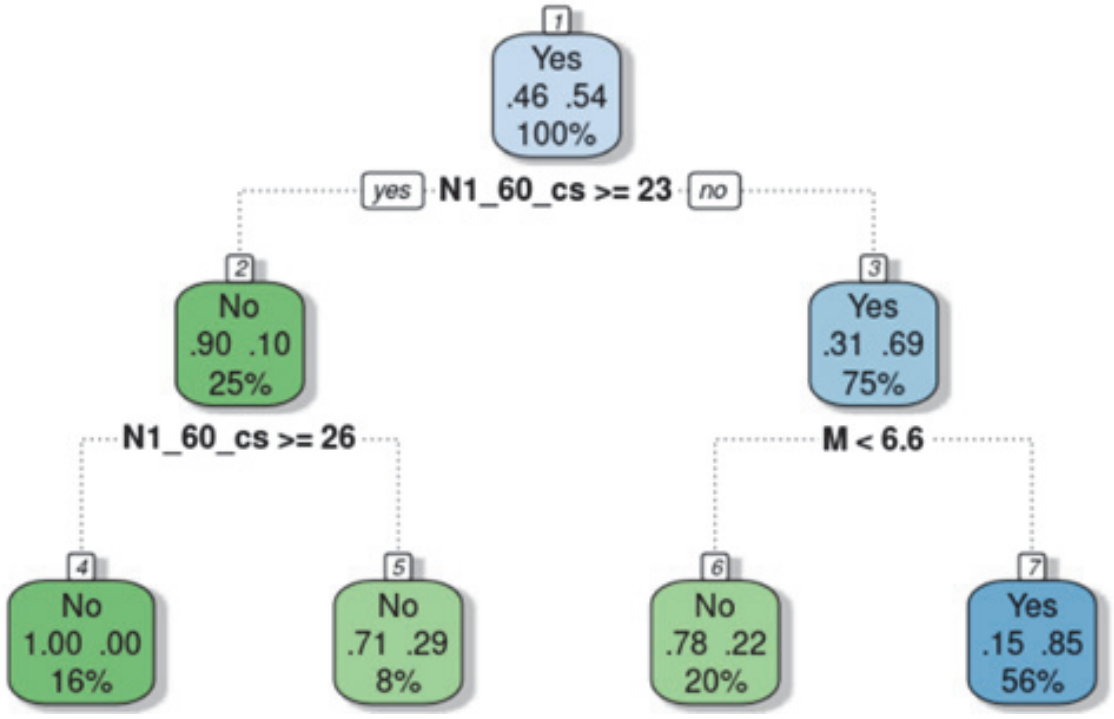


Fig. 17 – Árbol de decisión para estimar el potencial de licuefacción (base de datos en Cetin et al., 2018)

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

Intuitivamente no parece un método muy adecuado en tanto que no valora ni la continuidad ni la correlación espacial para un modelo del terreno bidimensional. No obstante, se ha demostrado preciso como modelo predictivo usando, como en este caso, las variables espaciales (x, y, z) como variables predictoras (o cota y distancia en el caso de la Figura 18).

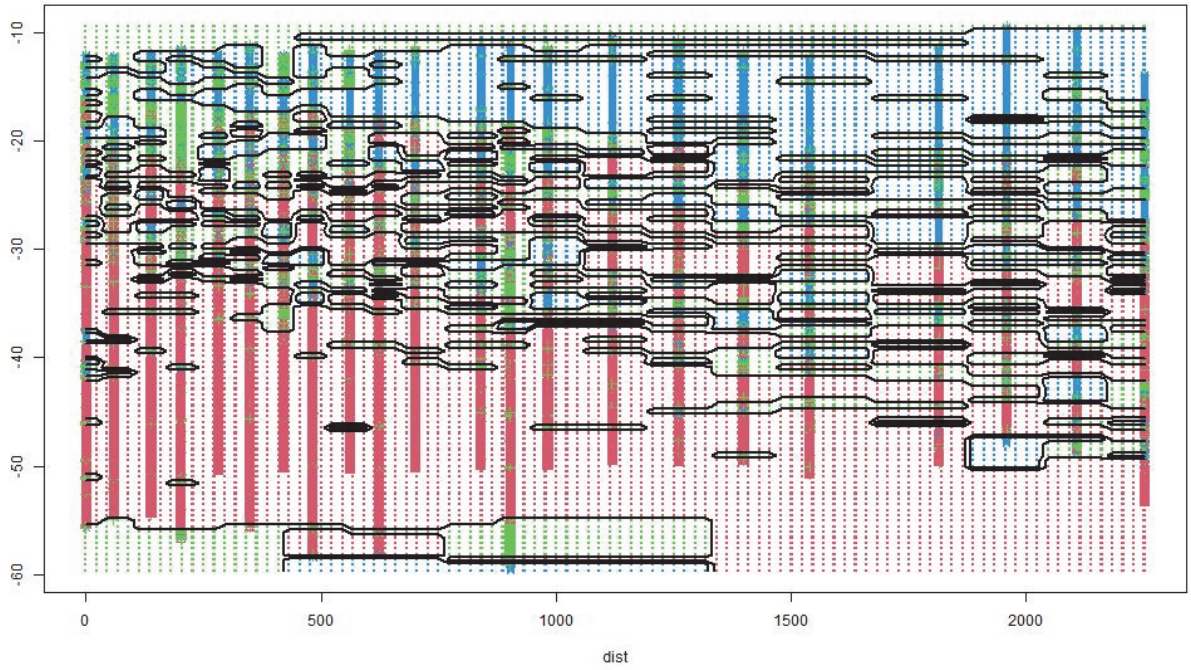


Fig. 18 – Modelo de terreno obtenido con el algoritmo de bosques aleatorios (dos variables predictoras: distancia y cota).

El modelo se expone con la voluntad de mostrar un ejemplo de perfil sobreajustado y poco práctico para el trabajo del proyectista, pero con buena capacidad predictiva. La técnica, con todo, tiene un campo de aplicación potencialmente amplio en el campo de la geotecnia.

5.2.6 – Redes neuronales

Posiblemente las redes neuronales representan el método más vanguardista y el estado del conocimiento del aprendizaje automático. El algoritmo está basado en una idea no excesivamente moderna, pero forman el corpus de lo que hoy se entiende, de forma tal vez, algo ostentosa, como "aprendizaje profundo" (Goodfellow et al (2016), Peña (2004), etc.). Las redes neuronales son, por tanto, la herramienta más potente, versátil, que mejor explota la capacidad computacional de los equipos actuales y que más representatividad tiene de lo que popularmente se entiende por técnicas aprendizaje automático.

Su popularidad se disparó en torno al año 2012, cuando la universidad de Toronto (Krizhevsky et al, 2012) publicó los resultados de una red neuronal "convolucional" diseñada para el reconocimento de imagenes, con el que consiguió una precisión del 85%. El diseño de este tipo de algoritmos está inspirado en los estudios del comportamiento del cerebro en el sistema visual de los animales. El hecho de que se haya desarrollado con tanto ímpetu en los últimos tiempos es debido a que las redes convolucionales precisan ser entrenadas con una gran cantidad de datos y, hasta hace poco, no se disponían de una capacidad de cálculo suficiente para poder entrenar una red multicapa en un tiempo razonable. Para clasificaciones no lineales de elevadas dimensiones, generar un modelo no lineal que se ajuste suficientemente a una solución, puede implicar un número de combinaciones de variables muy elevado y aproximar la solución puede ser impracticable.

Su justificación y funcionamiento proviene del teorema de Kolmogorov (Kolmogorov, 1957), según el cual cualquier función continúa de varias variables puede aproximarse como suma de funciones de una variable. Es decir, dada una función:

$$y = f(x_1, x_2, .., x_n) \tag{8}$$

Si f es continua y las variables x_i se encuentran entre 0 y 1 inclusive, se puede aproximar con tanta precisión como se desee mediante la combinación de funciones continuas de una variable (z) y N puede ser tan alto como se considere:

$$y = \sum_{i=1}^{N} g(z) \tag{9}$$

Las redes neuronales representan estrategias iterativas para resolver este problema.

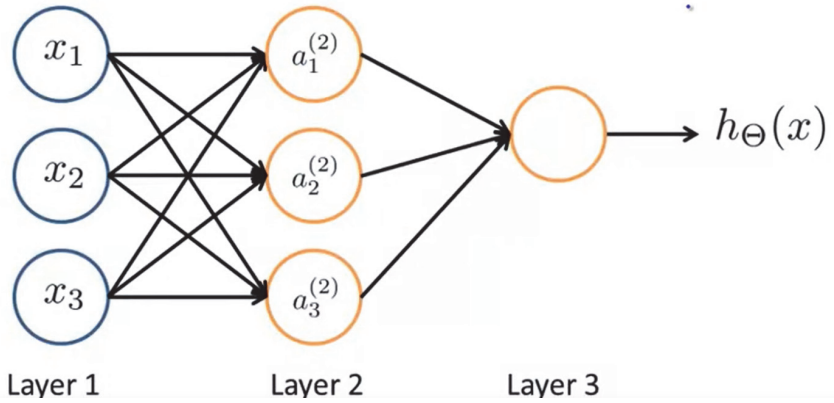


Fig. 19 – Modelo esquemático de redes neuronales.

Existen muchas estructuras de redes neuronales siendo la más utilizada para clasificación el "perceptrón" (Figura 19). Cada "neurona" de la capa intermedia se forma como una combinación lineal de las variables de entrada ($x = \{x_i\}$) transformando la combinación mediante una función g, tal que

$$a = g(w' \cdot x) \tag{10}$$

donde w es el vector de pesos, que puede o no incluir un término independiente, y g es la función de transformación no lineal. La función más habitualmente utilizada es la función logística y es la usada en el modelo representado en la Figura 20.

La literatura sobre el algoritmo es extensa y se puede encontrar sin dificultad. La obtención de la solución tiene un significativo número de matices y singularidades que conviene estudiar independientemente.

La aproximación realizada aquí, usando las variables espaciales como variables explicativas, fue propuesta, también, por el profesor Pijush Samui en su webinar sobre aprendizaje automático aplicado a la geotecnia (*https://www.issmge.org/education/recorded-webinars/machine-learning-in-geotechnical-engineering*). Aunque las redes neuronales sean uno de los algoritmos más flexibles y versátiles que existen, el estudio de la estructura óptima de la red (fundamentalmente el número de niveles ocultos) es difícil de predefinir y pudiera ser una herramienta poco práctica para formar parte del trabajo diario del geotécnico.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

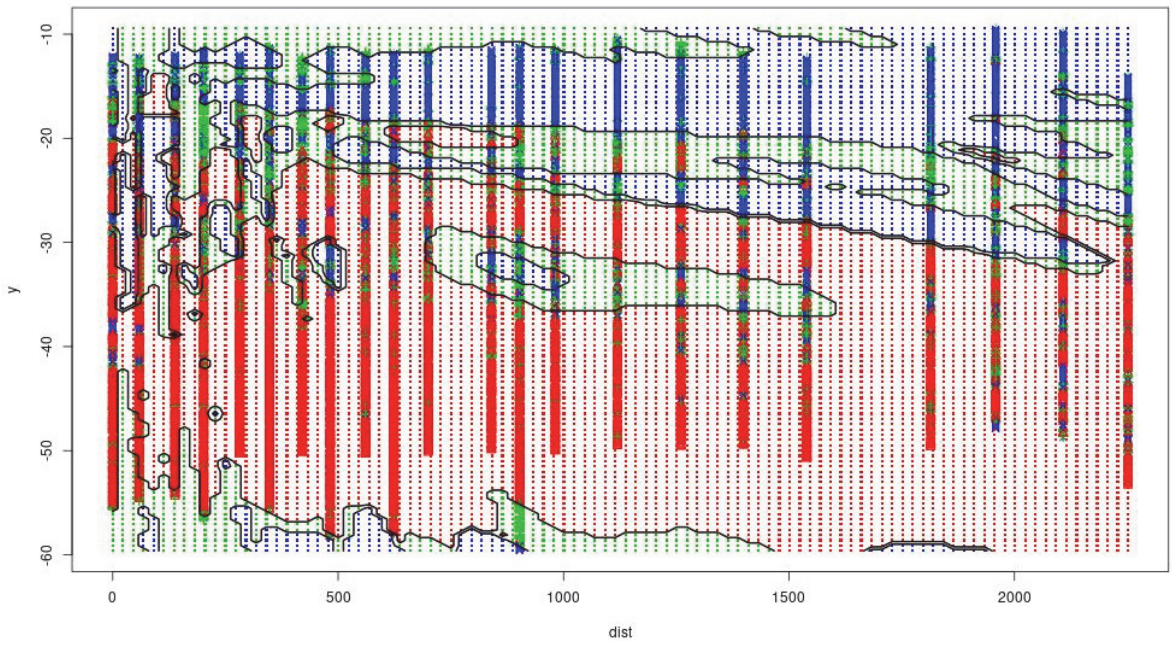


Fig. 20 – Modelo de terreno con redes neuronales; una capa oculta; 95 neuronas; decay=0.01.

Debido al elevado coste computacional de este algoritmo, se ha realizado una aproximación a la solución resultante con una única capa intermedia. Se han representado hasta 100 nodos en la capa oculta. Hasta donde se ha trabajado con este algoritmo, se ha comprobado que conduce a modelos razonables en el caso de estructuras simples del terreno, pero geológicamente rechazables en el caso de estructuras laminadas o más complejas.

5.2.7 – Máquinas de vector soporte (MVS)

Los algoritmos de máquinas de vector soporte (MVS) son otra de las técnicas más potentes y ampliamente utilizadas en aprendizaje automático y ofrecen una amplia gama de posibilidades y, en ocasiones, da resultados más "limpios" o, en el caso presente, visualmente más acordes a lo que se espera de un perfil geotécnico. Son una clase de modelos desarrollados originalmente por Vladimir Vapnik, a mediados de los 90, precisamente en el contexto de clasificación y busca establecer la frontera o límite más apropiado entre grupos de datos diferenciados lo que, en términos geotécnicos pudiera equipararse, informalmente, a lo que se hace al confeccionar un perfil del terreno a partir de un conjunto de datos clasificados.

El criterio que estableció Vapnik para definir la óptima de entre todas las posibles "rectas" de separación entre grupos de datos fue el de maximizar el "margen" entre grupos. El concepto de "margen" se ilustra en la Figura 21. De todas las fronteras que pudieran clasificar perfectamente los dos grupos, se da preferencia a la que maximiza la distancia entre ésta y el punto más cercano (Vapnik, 1999).

Esta separación no puede realizarse siempre mediante una línea recta. Al contrario que la mayoría de los métodos, que busca simplificar la información reduciendo el número de dimensiones del conjunto original y, sobre el conjunto reducido, busca la optimización, en el caso de este algoritmo, la estrategia es buscar un espacio de mayores dimensiones que permita formular una separación lineal. Mientras que los métodos clásicos buscan proyectar la información sobre un espacio de dimensión menor al original y, sobre este, establecer una separación no lineal; este criterio transforma los datos a un espacio de dimensión mayor que permita la separación lineal (por un hiperplano). El ejemplo que habitualmente se expone es aquel donde un grupo de datos no separable en una dimensión, pero sí lo es en dos dimensiones mediante una transformación que

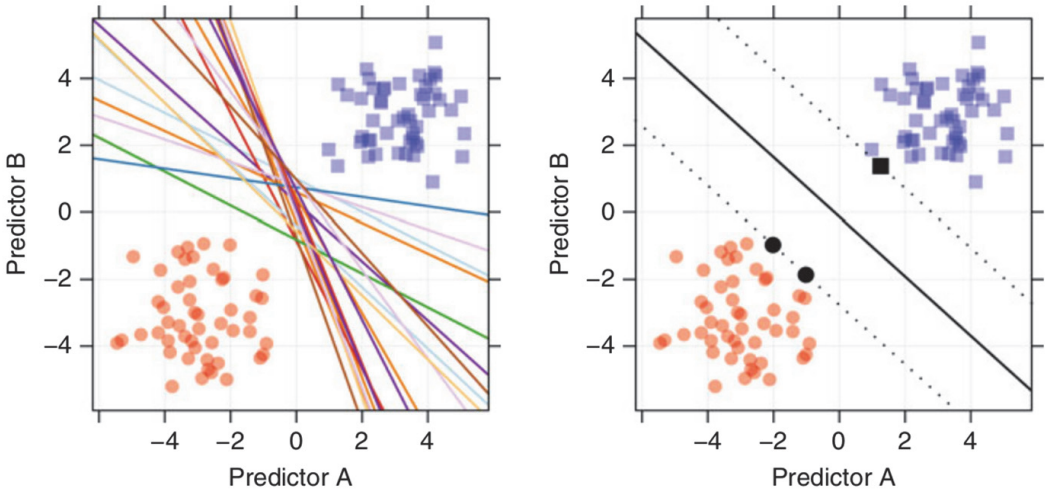


Fig. 21 – Concepto de "margen" para MVS; (izq.) dos grupos de datos diferenciados y algunas posibles fronteras que los clasificarían correctamente; (der.) clasificación lineal de máximo margen (Kuhn, 2016).

"eleva" en la dimensión (Figura 22). Al deshacer la transformación quedarían definidas las fronteras entre grupos.

Hay diversos parámetros de ajuste en el modelo de MVS. Se puede consultar detalles sobre sus implicaciones en la bibliografía (Hastie, T. et al (2008), Khun, M. y Jhonson , K (2016)), Abe, S. (2010), etc.):

- El kernel es la función de transformación y se puede entender igualmente como una medida o función de la "proximidad" o "similitud" entre puntos del espacio. En el caso presente se valoraron diversas opciones. El modelo mostrado se calculó con la función de base radial.
- El parámetro sigma (para el kernel de base radial), tiene el efecto de asignar cuándo un punto se encuentra más o menos próximo a otro.
- El parámetro de regularización C, tiene el efecto de evitar sobreajustes y previene de tender a modelos demasiado complejos con mala capacidad para generalizar. De forma similar a los modelos de regresión, se aplica como un sumando adicional de la función de coste.

Para visualizar cómo afectan C y sigma, se muestran en la Figura 23 los resultados para 3 valores de cada uno. Para el modelo descrito se ha realizado una validación cruzada por punto de investigación con 4 subdivisiones (4-fold).

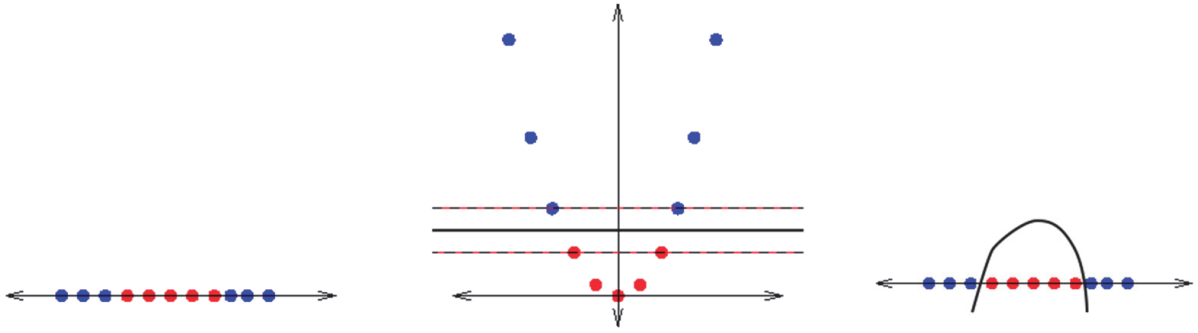


Fig. 22 – Puntos del espacio unidimensional linealmente no separables (izquierda), pero separables en dos dimensiones (centro). Al deshacer la transformación se tiene una definición de la frontera.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

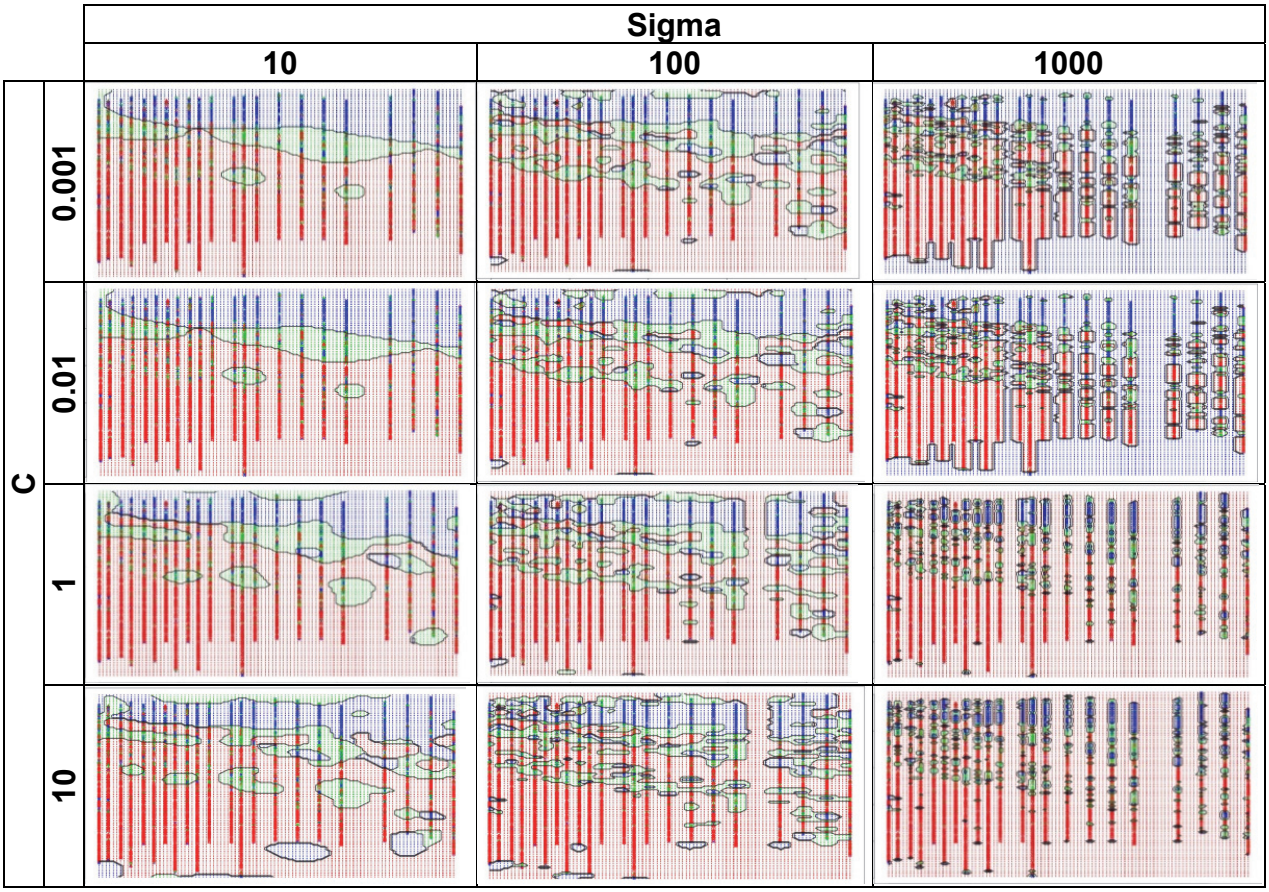


Fig. 23 – Variación de la definición entre fronteras para distintos valores de C y sigma.

De forma similar a como ocurría en el algoritmo de k-vecinos más próximos, la validación cruzada por punto (por CPTU) o aleatoria (tomando grupos de registros de forma aleatoria), resulta en evoluciones divergentes del parámetro kappa (Figura 24).

De nuevo, la validación cruzada por punto sirve de apoyo para el desarrollo de un modelo geológico práctico, resultando un indicador útil para determinar la generalidad que se está dispuesto a perder frente a su complejidad. El modelo adoptado se encuentra en la zona de

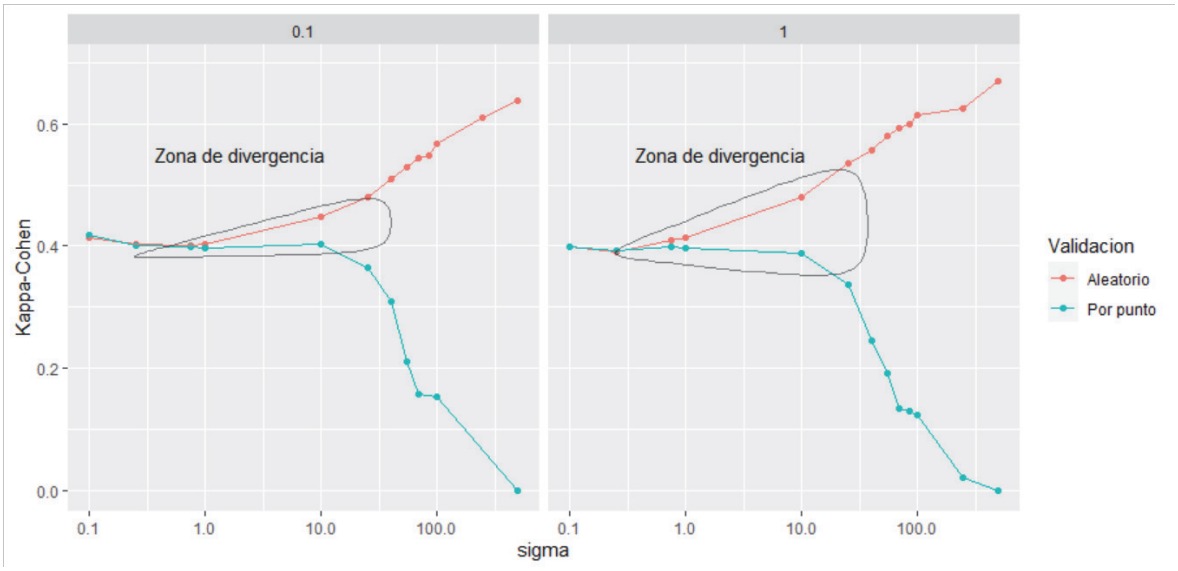


Fig. 24 – Resultados del modelo para validación cruzada por punto de investigación o aleatoria, para valores de C [0.1, 1] y de sigma [0.1, 1000].

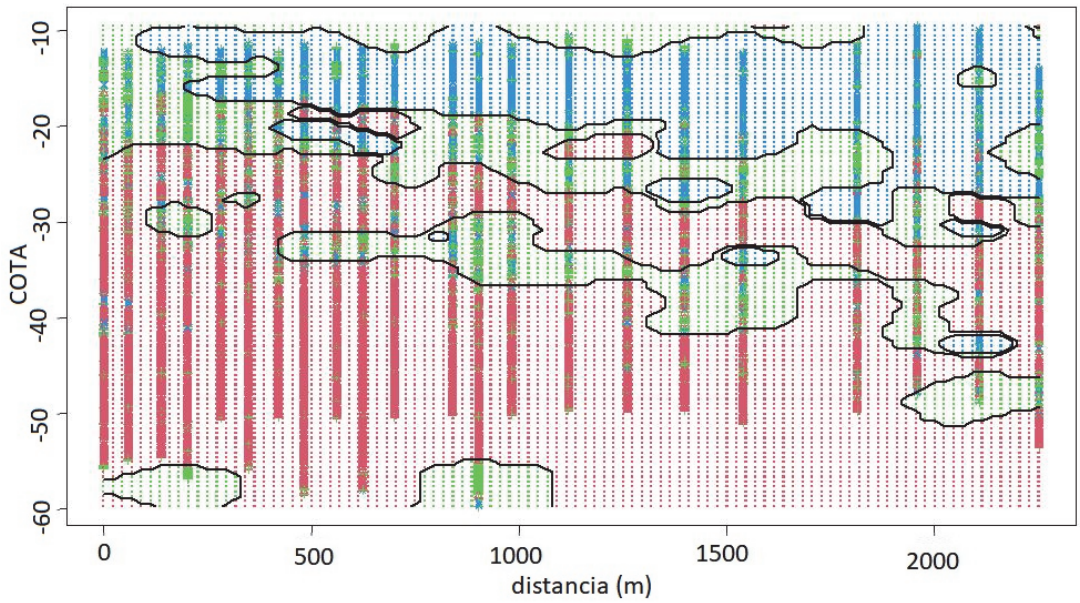


Fig. 25 – Resultados del modelo para valores de C=0.1 y sigma=40.

divergencia de las dos curvas y donde la validación por punto desciende de forma más abrupta. Como se ha dicho ya, y aunque dentro de ese rango de valores los modelos son sensiblemente iguales, debe ser el diseñador quien decida el grado de simplicidad del modelo.

5.2.8 – Modelo a aplicar

Se omiten aquí los otros modelos analizados en este estudio. En cualquier caso, la elección del tipo de algoritmo o modelo a utilizar no es evidente, puede entrañar dificultades y, en rigor, debe, ser dependiente de la morfología del terreno. En ocasiones conviene sacrificar la interpretabilidad (*k*-vecinos más próximo,…) de los resultados en favor de la flexibilidad y adaptabilidad del modelo (redes neuronales, bosques aleatorios, MVS). Un criterio razonable puede ser utilizar uno o varios modelos flexibles y que den resultados empíricos reconocidamente precisos y, posteriormente, valorar la posibilidad de adoptar un modelo más simple que llegue a resultados aproximadamente similares a los del modelo más complejo.

En este sentido, se ha comprobado que los modelos de *k*-vecinos más próximos y MVS dan lugar a resultados muy similares en el punto de inflexión del valor de *kappa* (y otros evaluadores) cuando se emplea la validación cruzada por punto de investigación, resultando perfiles del terreno satisfactorios desde el punto de vista geotécnico.

Independientemente de lo anterior, se entiende que la decisión del grado de simplicidad más apropiado para las condiciones específicas de la obra en estudio es tarea del diseñador geotécnico, amparándose en el valor de *kappa* o bien en la matriz de confusión en el caso de modelos predictivos.

Aunque no se ha mencionado de forma expresa, cualquiera que sea el perfil geotécnico final, resulta evidente que habrá que suprimir de la malla rectangular todos aquellos puntos que estén por encima del lecho marino y aquellos por debajo de las profundidades de exploración alcanzadas por los CPTUs (Figura 26).

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

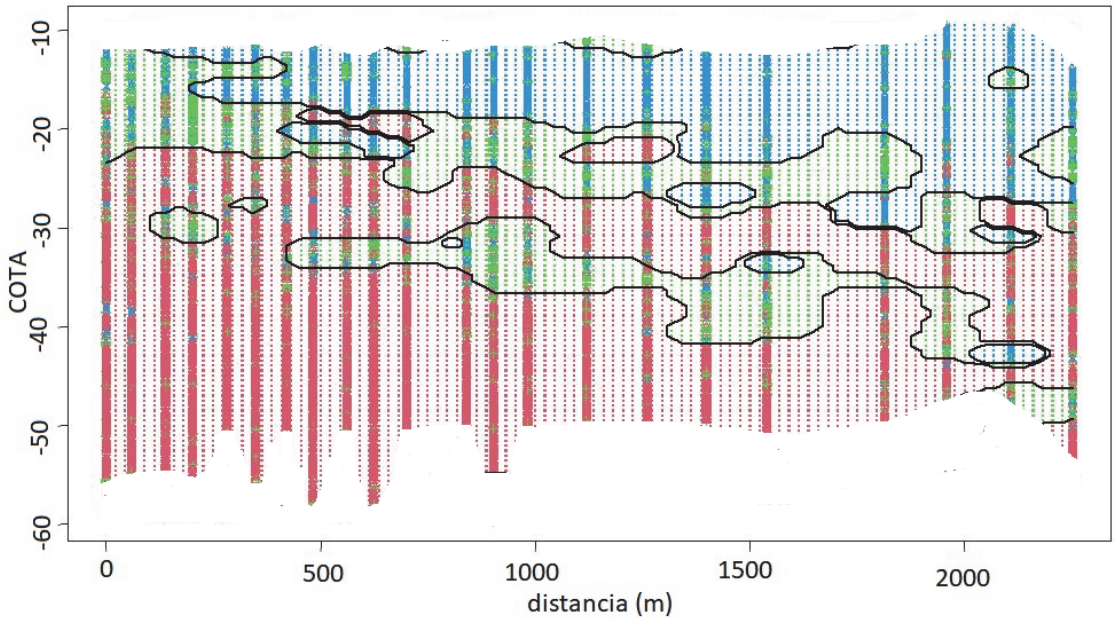


Fig. 26 – Resultado de un modelo tras suprimir puntos superficiales y profundos no explorados.

6 – CONSIDERACIONES FINALES

Este tipo de procesamiento racional de la información de los ensayos cumple una doble función: por un lado, puede ser usada de forma no ambigua, repetible y falsable para la elaboración de modelos geotécnicos; y por otro, puede ser utilizada como herramienta predictiva en campañas de campo en curso a fin de evaluar parcialmente la necesidad de información adicional de forma rápida y, también, sensiblemente objetiva (si bien no se han entrado directamente en la valoración de este tipo de aproximaciones en este artículo.

Estas técnicas aportan diversas ventajas respecto a la aproximación clásica. Por un lado, son fáciles de implementar y se integran de forma natural en la filosofía de trabajo BIM, permitiendo la automatización y optimización de la fase de interpretación geotécnica o geológica. Por otro, permiten, al menos parcialmente, objetivar y evaluar numéricamente, la elaboración de un perfil geotécnico, tarea históricamente algo subjetiva. En cualquier caso, estas aproximaciones son aún muy preliminares y requieren acumular experiencia para cualquier aplicación inmediata. No se ve una perspectiva próxima en la que no se considere absolutamente imprescindible la supervisión y validación del modelo por parte de un geólogo experimentado. Este tipo de aproximación en la ejecución de perfiles permite (en el caso de ser viable), que el proyectista decida en qué medida simplificar el modelo pudiendo evaluar objetivamente esta simplificación.

Para la elaboración de los modelos entran en juego muy diversas variables y, su influencia en ellos solo se han evaluado aquí de forma somera: el tipo de algoritmo a utilizar, medida de distancia, tipo de validación cruzada, tipo de regularización, centrado y escalado (o no) de valores de entrada, uso (o no) de distancias logarítmicas para algunas de las variables, tipo de medida de evaluación del modelo, parámetros específicos de cada modelo, etc. Su validez, por tanto, se reduce a los perfiles estudiados.

El coste computacional para la elaboración de los modelos depende claramente del tamaño muestral, del modelo y del tipo de validación. Así, en campañas con pocos CPTUs es probable que se puedan aplicar sin problemas muchos modelos; sin embargo, en grandes campañas de CPTUs a grandes profundidades puede resultar inviable la elaboración de un modelo con todas las lecturas de campo. Se ha demostrado que el uso de submuestras aleatorias de registros tan continuos como los del CPTU altera poco el resultado del modelo.

7 – CONCLUSIONES

Se han aplicado diversas técnicas de aprendizaje automático con los ensayos CPTU de una campaña portuaria para la elaboración de perfiles geotécnicos.

Para su obtención se ha procedido de manera secuencial con un algoritmo de aprendizaje no supervisado para la categorización de los registros del ensayo (Q_{tn}, B_q y Fr(%)), y un algoritmo de aprendizaje supervisado (sin grupo de validación, actuando como no supervisado) para la elaboración de un modelo del terreno basado en la categorización anterior.

Se ha comprobado que el algoritmo de máquinas de vector soporte, en este contexto, proporciona modelos útiles y que puede implementarse de forma rápida y automatizada. Igualmente, el algoritmo de k-vecinos más próximos proporciona resultados similares, que pueden utilizarse de forma similar con un coste computacional claramente menor. Todos los algoritmos de agrupamiento (no supervisados) utilizados en este contexto aportan resultados similares e igualmente prácticos.

Para la elaboración de los modelos se requiere realizar una validación cruzada extrayendo registros completos de uno o varios puntos de investigación en lugar de registros aleatorios. Operando así, se puede estudiar el punto de inflexión que define el modelo y que, a juicio de los autores, resulta un indicador razonable para la elaboración de un modelo geológico práctico. En caso de querer elaborar modelos predictivos más precisos, la validación cruzada aleatoria tiende a dar resultados significativamente mejores.

8 – AGRADECIMIENTOS

Los autores desean mostrar su agradecimiento a Puertos del Estado por el continuo apoyo en este y otros proyectos de investigación desarrollados por el CEDEX y por su evidente compromiso por la excelencia técnica.

También desean agradecer y resaltar el compromiso que la Autoridad Portuaria de Barcelona ha mostrado por una investigación geotécnica de calidad en el marco de los proyectos de desarrollo del Puerto de Barcelona, así como su disposición a cualquier propuesta de desarrollo técnico orientada a la profundización en la geotécnica y conocimiento del terreno.

Al equipo de campo de Igeotest por su profesionalidad y calidad en los trabajos desarrollados en el entorno marino.

9 – REFERENCIAS BIBLIOGRÁFICAS

Abe, S. (2010). *Support Vector Machines for Pattern Classification*. 2ª Edicion. Springer.

Cetin, K.; Seed, R.; Kayen, R.; Moss, R.; Bilge, T.; Ilgaç, M.; Chowdhury, K. (2018). *Dataset on SPT-based seismic soil liquefaction. Data in Brief*. 20. 10.1016/j.dib.2018.08.043.

Duda, R. O.; Hart, P. E.; Stork, D. G. (2000). *Pattern Classification*, 2ª edición, John Wiley & Sons Inc

EN 1997-2 (2007): *Eurocode 7: Geotechnical design - Part 2: Ground investigation and testing*. CEN, Brussels, Belgium.

Hartigan, J.A.; Wong, M. (1979). *A k-means clustering algorithm*. Applied Statistics 28, 100–108.

Hastie, T.; Tibshirani, R.; Friedman, J. (2008). *The elements of statistical learning: Data mining, inference and prediction*. 2ª edición. Springer.

Hegazy, Y.A.; Mayne, P.W. (2002). *Objective site characterization using clustering of piezocone data*, Journal of Geotechnical and Geoenvironmental Engineering, 128(12): 986–996.

ISSN 0379-9522; e-ISSN 2184-8394 – Geotecnia nº 152 – julho/julio/july 2021 – pp. 585-615
https://doi.org/10.14195/2184-8394_152_19 – © 2021 Sociedade Portuguesa de Geotecnia

Khun, M.; Johnson, K. (2016). *Applied predictive modelling*. 5ª Edición. Springer.

Kolmogorov, A. N. (1957). *On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition.* Dokl. Akad. Nauk SSSR, 114:5 (1957), 953–956.

Krizhevsky, A.; Sutskever, I.; Hinton E.H. (2012). *ImageNet Classification with Deep Convolutional Neural Networks.* Neural Information Processing Systems. 25. 10.1145/3065386

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

López Mántaras, R. (2020). *El traje nuevo de la inteligencia artificial.* Investigación y Ciencia, 526, pp 52-59.

MacQueen, J.B. (1967). *Some methods for classification and analysis of multivariate observations*. In: Le cam, L.M., Neyman, J. (Eds.), Proceedings of 5th Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 281–297.

Mlynarek, Z.; Wierzbicki, J.; Wolynski, W. (2005). *Use of cluster method for in situ tests*. Studia Geotechnica et Mechanica, XXVII(3–4): 15–27.

Murphy, K. P. (2012). *Machine Learning: A probabilistic perspective*. Massachusetts Institute of Technology.

Nadim, F. (2007). *Tools and Strategies for Dealing with Uncertainty in Geotechnics*. In: Griffiths D.V., Fenton G.A. (eds). Probabilistic Methods in Geotechnical Engineering. Springer.

Peña, D. (2004). *Análisis de datos multivariantes*. McGraw-Hill Interamericana de España S.L.

Robertson, P.K. (2016). *Cone penetration test (CPT)-Based soil behaviour type (SBT) Classification system- an update*. Canadian geotechnical journal, 53, pp. 1910-1927.

Vapnik, V. N. (1999). *The nature of statistical learning theory.* 2ª edición. Springer.

Wierzbicki J.; Smaga A.; Stefaniak K.; Wołyński W. (2016). *3D mapping of organic layers by means of CPTU and statistical data analysis*. Geotechnical and Geophysical Site Characterisation 5 – Lehane, Acosta-Martínez & Kelly (Eds.)

Wierzchon, S. T.; Klopotek, M. A. (2018). *Modern algorithms of cluster analysis*. Springer.