7º CONGRESO FORESTAL ESPAÑOL

Gestión del monte: servicios ambientales y bioeconomía

26 - 30 junio 2017 | Plasencia
Cáceres, Extremadura

7CFE01-136

Edita: Sociedad Española de Ciencias Forestales
Plasencia. Cáceres, Extremadura. 26-30 junio 2017
ISBN 978-84-941695-2-6

Sociedad Española de Ciencias Forestales
SECF

Clasificación de la calidad de estación forestal mediante técnicas de aprendizaje automático (*machine learning*)

BRAVO OVIEDO, F.[1,2] , BRAVO NÚÑEZ, A.[3]

[1] Instituto Universitario de Investigación y Gestión Forestal Sostenible. Universidad de Valladolid - INIA. Avda.Madrid s/n. 34004. Palencia. España. fbravo@pvs.uva.es

[2] Departamento de Producción Vegetal y Recursos Forestales, ETS de Ingenierías Agrarias. Universidad de Valladolid. Palencia, España.

[3] Grado INDAT (Informática + Estadística) Universidad de Valladolid andres.bravo@alumnos.uva.es

Resumen

La estimación de la calidad de estación es una de las claves de la modelización de la producción forestal. Dado que la productividad de las estaciones forestales está cambiando es clave disponer de sistemas adaptativos que permitan predecir de forma dinámica la calidad de estación. Las técnicas de aprendizaje automático (conocidas en inglés como *machine learning*) permiten que la clasificación de entes (en nuestro caso las estaciones forestales) se haga de forma dinámica. A partir de un conjunto de datos de *Pinus sylvestris* L. en el Alto Valle del Ebro, se han desarrollado procedimientos de aprendizaje supervisado para clasificar la calidad de estación mediante regresión logística multinomial, análisis discriminante lineal y el procedimiento de Bayes ingenuo (*naive Bayes*) a partir de datos edáficos. El modelo seleccionado se base en el procedimiento Bayes ingenuo y clasifica correctamente, mediante un procedimiento de validación cruzada, el 64,3% de las observaciones a partir de la concentración de potasio (transformada mediante las raíz cuadrada) y el porcentaje de materia orgánica.

Palabras clave

Regresión logística multinomial, Bayes, Análisis discriminante, *Pinus sylvestris*, Productividad

1. Introducción

Los modelos de crecimiento y producción forestal dependen en gran medida de una clasificación adecuada de la productividad de la estación. Las curvas de calidad en las que a partir de la altura dominante (H_0) y la edad del rodal puede estimarse la calidad de estación son el método habitual para calcular el índice de sitio (Skovsgaard y Vanclay, 2007) Sin embargo este método asume la estabilidad en la productividad de la estación forestal y la disponibilidad de árboles adecuados para medir la altura dominante. En los rodales donde se han aplicado cortas por dimensión (huroneo o floreo) no se puede asegurar una correcta medición de la altura dominante por falta de árboles que hayan pertenecido al estrato dominante durante toda su vida (Bravo y Montero, 2001) Además, la estabilidad de la calidad de estación en el tiempo también está considerada como no adecuada (Spiecker et al, 1996) lo que es cada vez más plausible dada el cambio climático que estamos observando (IPCC, 2014) Por tanto cuando se quiere predecir la calidad de estación a largo plazo no debieran emplearse las curvas de calidad sino métodos de clasificación basados en variables ambientales. Estos métodos de clasificación a partir de datos ambientales han sido desarrollados en los últimos años para diversos tipos de masas forestales en el mediterráneo occidentela como pinares naturales de *Pinus sylvestris* L. en el Alto Valle del Ebro (Bravo y Montero, 2001) y de *Pinus pinea* L. en Huelva (Bravo-Oviedo y Montero, 2005), repoblaciones de *Pinus pinea* L. en Calabria, Italia (Bravo et al, 2011), Pinus sylvestris en los páramos ácidos de Castilla y León (Bueis et al, 2016) o de *Pinus radiata* Ait. en Galicia (Sánchez-Rodríguez et al, 2002) o alcornocales (*Quercus suber* L.) en Portugal (Paulo et al, 2015).

Las aproximaciones estadísticas utilizadas en estos estudios para clasificar las calidades de estación ha sido fundamentalmente el análisis discriminante lineal (Bravo y Montero, 2001, Bravo-Oviedo y Montero, 2005, Bravo et al, 2011, Bueis et al, 2016) pero también se ha utilizado la regresión lineal múltiple (Sánchez-Rodríguez et al, 2002) o la regresión por mínimos cuadrados parciales (Paulo et al, 2016).

Sin embargo, recientemente se ha producido un incremento notable del uso de los métodos de clasificación en el campo de la inteligencia artificial para la identificación de patrones (p.ej., mercadotecnia, detección de *spam* en el correo electrónico, fraudes bancarios, detección de enfermedades tanto en humanos como en animales,...) Entre estos métodos de clasificación destaca el uso del Bayes ingenuo o *Naive Bayes*, la regresión logística multinomial y el análisis discriminante lineal.

2. Objetivos

El objetivo de este trabajo es determinar cuál es el método de aprendizaje automático más adecuado para la clasificación de rodales de *Pinus sylvestris* L. en el Alto Valle del Ebro en calidades de estación a partir de datos edáficos. Para cumplir este objetivo se han desarrollado 127 modelos diferentes mediante tres métodos diferentes de clasificación: Bayes ingenuo, regresión logística multinomial y análisis discriminante lineal.

3. Metodología

Área de estudio

Los datos fueron obtenidos en pinares de *Pinus sylvestris* L en la zona de transición conocida como Alto Valle del Ebro entre las comarcas del Valle de Losa (Burgos) y de Valdegovía/Gaubea (Álava/Araba) situados entre 700 y 900 metros sobre el nivel del mar. El clima predominante es una transición entre los tipos mediterráneo y atlántico sin sequías ni heladas intensas y con una precipitación anual promedio de 787 mm de los que 123 mm corresponden a precipitaciones estivales y una temperatura promedio anual de 11,2 ºC. Lo suelos son calcáreos con profundidad entre 10 y 60 cm y pH de 6,5 a 8,4

Selvicultura aplicada a los pinares estudiados

Hasta principios del siglo XX el método de ordenación aplicado ha sido el de tramos permanentes y las cortas de regeneración por aclareos sucesivos y uniformes con turno de 88 años y periodo de regeneración de 22 años, se mantuvo la aplicación de las cortas por dimensión (huroneos o floreos)

Datos

Se han utilizado 28 parcelas procedentes de un estudio previo sobre calidad de estación (curvas de calidad y regla discriminante) en las masas objeto de estudio (Bravo y Montero, 2001) En cada una de las parcelas se determinó el índice de sitio (IS), definido como la altura dominante a los 100 años, a partir de las curvas de calidad desarrolladas por Bravo y Montero (2001) Las 28 parcelas representan las cuatro calidades definidas: IS 14 metros (15 parcelas), 17 metros (4 parcelas), 20 metros (5 parcelas) y 23 metros (4 parcelas) Se utilizaron además los datos edáficos (tabla 1) obtenidos a partir de una muestra de los primeros 10 cm de acuerdo con lo propuesto por Jokela et al (1988) Las muestras de suelos se analizaron para determinar su textura (porcentajes de arena, arcilla y limo) siguiendo el método propuesto la Sociedad Internacional de Ciencias del Suelo (ISSS) Además se determinaron los porcentajes de carbonatos, caliza activa y materia orgánica; fósforo en partes por millón (ppm) según el método Olsen, potasio en ppm utilizando acetato amónico 1 N, calcio, magnesio y sodio en meq/100 g utilizando también acetato amónico 1 N, Capacidad de Intercambio Catiónico (CCC), pH y conductividad en mmhos/cm.

Tabla 1. Número de muestras, valores medios, máximo y mínimo y desviación típica de las variables edáficas medidas en las parcelas en que se ha podido determinar el SI.

VARIABLE	Media	Desv. típica	Máximo	Mínimo
Arena (%)	62,08	14,06	85,23	37,05
Limo (%)	20,06	9,13	37,40	6,75
Arcilla (%)	16,80	8,67	34,85	5,35
P (ppm)	3,50	4,48	23,00	1,00
K (ppm)	140,32	97,59	445,00	19,00
Ca (meq/100g)	16,91	13,69	49,40	0,20
Mg (meq/100g)	0,19	0,21	0,60	0,00
Na (meq/100g)	0,84	1,08	4,95	0,05
Carbonatos (%)	3,37	5,49	18,00	0,00
Materia org. (%)	4,89	2,06	10,30	1,10
CCC (meq/100g)	19,16	8,56	34,40	5,50
pH	6,26	1,57	8,30	3,60
Conductividad (mmhos/cm)	0,19	0,11	0,41	0,02

Análisis estadístico: normalidad y correlación

Se ha tomado como base el análisis de normalidad (mediante el test de Shapiro-Wilk), tanto de variables originales como transformadas mediante la raíz cuadrada, y correlación (mediante el coeficiente de Pearson) entre variables explicativas realizado por Bravo y Montero (2001) por lo que se ha partido de solo cinco variables candidatas (tablas 2 y 3) para los diferentes métodos de clasificación ensayados.

Tabla 2. Probabilidades de aceptación de la hipótesis nula del test de normalidad de Shapiro-Wills. En negrita se indican los casos en que la variable se distribuye normalmente.

VARIABLE	X	√(X)
Arena	**0,3115**	0,2853
Limo	**0,1909**	0,3064
Arcilla	0,0436	**0,3055**
P	0,0001	0,0001
K	0,0058	**0,7296**
Ca	0,0038	0,0488
Mg	0,0001	0,0001
Na	0,0001	0,0040
Carbonatos	0,0001	0,0001
Materia orgánica.	**0,5217**	0,5211
CCC	**0,1117**	0,0354
PH	0,0048	0,0028
Conductividad	**0,2831**	0,4260

Tabla 3. Correlación entre las variables edáficas, normales o normalizadas.

Variables	Arena	Limo	√(Arcilla)	√(K)	Mat. org.	CCC	Conduct.
Arena	1,00000	-0,77622	-0,75655	-0,63352	-0,07854	-0,28235	-0,65050
Limo		1,00000	0,41586	0,67842	0,32249	0,48531	0,72451
√(Arcilla)			1,00000	0,53364	0,00948	0,11875	0,48101
√(K)				1,00000	0,33484	0,39082	0,53741
Materia orgánica					1,00000	0,49158	0,54512
CCC						1,00000	0,37980
Conductividad							1,00000

Análisis estadístico: métodos de clasificación

Para poder clasificar las observaciones en cada una de las cuatro clases de índice de sitio descritas se han utilizado tres métodos de clasificación: *naive* Bayes o Bayes ingenuo, análisis discriminante lineal y regresión logística multinomial. Todos los métodos de clasificación se han analizado con el paquete estadístico R (R Core Team, 2016) y los paquetes MASS (Venables y Ripley, 2002), e1071 (Meyer et al, 2015) y nnet (Venables y Ripley, 2002)

Bayes ingenuo (naive Bayes)

El método de clasificador Bayes ingenuo (Hastie et al, 2013 y en Hand y Yu, 2001 se puede obtener una revisión sobre este método) se basa en el teorema de Bayes y la independencia de las variables explicativas en nuestro caso fundamentada porque no se han introducido de forma conjunta variables que presentasen una alta correlación (tabla 3) La probabilidad de que una observación pertenezca a una clase determinada k (en nuestro caso a una calidad de estación) es igual a:

$$p(C_k|x_1,\dots,x_n) = \frac{p(C_k,x_1,\dots,x_n)}{\sum_{k=1}^{n} p(C_k,x_1,\dots,x_n)} = \frac{p(C_k,x_1,\dots,x_n)}{p(x_1,\dots,x_n)}$$

Donde C_k son las clases definidas (en nuestro caso 4 calidades de estación), x_i son las variables independientes y p son las probabilidades de ocurrencia. $p(C_k,x_1,\dots,x_n)$ es la probabilidad de ocurrencia, a priori, de cada clase (en nuestro caso la proporción de cada calidad de estación obtenida a partir del muestreo), $p(x_1,\dots,x_n)$ es constante para cada conjunto de datos y la probabilidad condicional de las k clases definidas , si consideramos independencia entre las variables, se calculará como:

$$p(C_k|x_1,\dots,x_n) = \frac{p(C_k)\prod_{i=1}^{n} p(x_i|C_k)}{p(x_1,\dots,x_n)}$$

La regla de clasificación en este caso es asignar cada observación a la clase con probabilidad condicional de asignación más alta, es decir a la más probable.

Análisis Discriminante Lineal

El método de análisis discriminante linear se basa en que la distribución de las variables es normal, con igual matriz de covarianzas entre grupos y independientes entre ellas (Hastie et al, 2013, James et al, 2015). Se buscan direcciones que maximicen el cociente de Rayleigh (dispersión entre grupos/dispersión dentro de un grupo) para ello se estima la dispersión dentro de cada grupo (S_w) mediante la siguiente expresión:

$$S_w = \sum_{i=1}^{c} \sum_{x \in C_i}^{n_i} (x - m_i) * (x - m_i)^T$$

Por otro lado la dispersión entre grupos (S_b) se estima mediante:

$$S_b = \sum_{i=1}^{c} n_i (m_i - m) * (m_i - m)^T$$

Donde m_i es la media que representa el centroide de cada grupo *i*, *c* es el número de grupos, n_i es el número I de observaciones *x* de cada grupo.

La dispersión total (T) es la suma de estas dos matrices (S_w+S_b). La dirección que maximiza el cociente es el autovalor correspondiente al mayor autovector de T^{-1}*B donde B es el valor real de la

dispersión entre grupos que se estima mediante la varianza muestral S_b. Se repite el proceso hasta obtener todos los ejes discriminantes. Después se proyectan los centroides de los grupos en el hiperplano definido. Así dada una nueva observación esta se proyecta en el hiperplano y se asigna al grupo cuyo centroide se encuentre más cerca utilizando (como norma general) la distancia de Mahalanobis.

Regresión logística multinomial

La regresión logística (Hastie et al, 2013) permite calcular la probabilidad de pertenencia a cada una de las k clases definidas mediante k-1 modelos logísticos que permiten cada uno determinar la probabilidad de pertenencia a una clase determinada y al restar a 1 la suma de todas ellas obtener la probabilidad de pertenencia a la última clase. Se puede expresar de esta forma:

$$p_j = \frac{1}{1 + e^{-Z}}$$

Donde p_j es la probabilidad de que una observación pertenezca a las clases 1,2,... k-1 y Z es una función lineal que se ajusta para cada k-1 clases definidas

$$Z = \beta_0 + \sum \beta_i x_i$$

La probabilidad de pertenencia a la clase k se calcula como $p_k = 1 - \sum p_j$

Análisis estadístico: selección de modelos y comparación

Para cada uno de los procedimientos descritos se ha seleccionado el mejor modelo de entre los 127 posibles (valor que se obtiene al combinar las siete variables de todas las formas posibles, desde una variable sola hasta las siete de forma conjunta) mediante la validación cruzada al clasificar cada observación con el modelo ajustado con todas las observaciones menos esa (*leave-one-out*) La proporción de aciertos globales (A_G), en todas las clases, se ha utilizado para seleccionar los mejores modelos dentro de cada procedimiento y posteriormente se ha refinado esta selección teniendo en cuenta las tasas de acierto por grupos (número de observaciones de una clase correctamente clasificadas partido por todas las observaciones de esa clase). Además se ha utilizado la precisión por clases que se calcula como el número de predicciones correctas (se clasifica como perteneciente a esa clase observaciones que pertenecen a la misma) de la clase dividida por el total de las predicciones en esa clase (tanto si pertenecen a ella como si no). Finalmente los modelos seleccionados para cada procedimiento se han comparado de la misma forma entre ellos y para el mejor modelo se ha calculado el coeficiente de Kappa (K) de Cohen (1960) definido en este caso a partir de la proporción de la calidad más frecuente (P_f):

$$K = \frac{A_G - P_f}{1 - P_f}$$

4. Resultados

Mediante las tres técnicas ensayadas (Bayes ingenuo, análisis discriminante lineal y regresión logística lineal) se han extraído los cuatro mejores modelos a partir del valor de aciertos globales (A_G) y en caso de empate entre modelos se han incluido todos con el mismo valor (tabla 4). La regresión logística multinomial presenta el mayor acierto global cuando se utilizan como variables explicativas la conductividad y la raíz cuadrada del porcentaje de arcilla (modelo 15 en la tabla 4). Sin embargo, para la calidad de estación 17 no predice ninguna observación y por tanto el modelo presenta una carencia grave ya que ningún rodal es clasificado en esa calidad. Los mejores modelos que no presentan este problema son los modelos 5, 8 y 13. De entre ellos el modelo 5, estimado mediante el

método de Bayes ingenuo, es el que presenta un mayor valor de aciertos globales (64,3%) y unas tasas de aciertos globales por calidades razonables (entre 0,25, para calidad 17, y 0,80 `para la calidad 14). Además la precisión por clases (tabla 3) oscila entre 1 para las calidades 17 y 23 y 0,333 para la calidad 20. Las variables explicativas del modelo 5 son la materia orgánica y la raíz cuadrada del potasio.

Tabla 4. Proporción de aciertos globales y por clase (A), precisión por clases (P) y variables explicativas para los cinco mejores modelo obtenidos para cada uno de los procedimientos. Los subíndices denotan los valores globales (G) y por Índice de Sitio (23, 20, 17 y 14 metros de altura dominante a los 100 años) na significa que no hay predicciones para esa clase y por tanto el denominador de la precisión es igual a cero.

Modelos	AG	A14	A17	A20	A23	P14	P17	P20	P23	Variables explicativas
Bayes Ingenuo (Naive Bayes)										
1	0,571	0,867	0,25	0,00	0,50	0,65	0,333	0,00	0,50	√(Arcilla), √(K)
2	0,571	0,80	0,5	0,00	0,50	0,75	0,667	0,00	0,50	Arena, √(Arcilla), √(K), Conductividad
3	0,607	0,933	0,00	0,00	0,75	0,609	na	0,00	0,75	√(Arcilla), Conductividad
4	0,607	0,933	0,25	0,00	0,50	0,636	1	0,00	0,50	√(Arcilla), √(K), Conductividad
5	0,643	0,80	0,25	0,40	0,75	0,667	1	0,333	1	√(K), Materia orgánica
Análisis Discriminante Lineal										
6	0,607	0,933	0,00	0,00	0,75	0583	na	na	0,75	Conductividad
7	0,607	0,867	0,25	0,00	0,75	0,591	0,50	na	0,75	√(Arcilla), Materia orgánica
8	0,607	0,80	0,50	0,00	0,75	0,632	0,667	0,00	0,60	√(Arcilla), Conductividad
9	0,643	1,00	0,00	0,00	0,75	0,6’	na	na	1,00	Materia orgánica
Regresión Logística Multinomial										
10	0,607	1,00	0,00	0,00	0,50	0,652	na	na	0,40	√(K)
11	0,607	0,933	0,00	0,00	0,75	0,657	0,60	0,00	0,60	Conductividad
12	0,607	0,733	0,75	0,00	0,75	0,647	0,60	0,00	0,60	√(Arcilla), √(K), Conductividad
13	0,607	0,733	0,75	0,00	0,75	0,647	0,60	0,00	0,60	√(Arcilla), CCC, Conductividad
14	0,643	1,00	0,00	0,00	0,75	0,60	na	na	1,00	Materia orgánica
15	0,679	0,867	0,75	0,00	0,75	0,684	0,75	na	0,60	√(Arcilla), Conductividad

A partir de la matriz de confusión (tabla 5) del modelo Bayes ingenuo a partir de la materia orgánica y la raíz cuadrada del potasio (modelo 5) se puede observar que la predicción de la clase de índice de sitio 17 solo se clasifica correctamente en un 25% de las ocasiones mientras que para la clase 14 lo hace correctamente un 80% de las ocasiones.

Tabla 5 Matriz de confusión del modelo 5, estimado mediante Bayes ingenuo para clasificar la calidad de estación de rodales de Pinus sylvestris L. en el Alto Valle del Ebro.

		Calidad de estación predicha			
		14	17	20	23
Calidad de estación real	14	12	0	3	0
	17	2	1	1	0
	20	3	0	2	0
	23	1	0	0	3

5. Discusión

La mayor parte de los métodos de clasificación funcionan mejor cuando el número de datos utilizados para su ajuste es alto. En la mayor parte de los trabajos de clasificación de la calidad de estación el número de observaciones disponibles es bajo, habitualmente entre 30 y 50 observaciones (p.ej., Bravo y Montero, 2001, Sánchez et al, 2002, Bravo-Oviedo et al, 2005 o Bueis et al, 2016) pero con notables excepciones como el trabajo de Paulo et al (2015) que se basa en 100 observaciones. En cualquier caso lejos de los valores habituales para estos métodos en otros campos científicos. En este trabajo la tasa de acierto global (64,3%) no está lejos de las obtenidas en trabajos similares que habitualmente está entre el 60 y el 75% (ver por ejemplo, Bravo y Montero, 2001, Bravo et al, 2011 o Bueis et al, 2016) pero con gran variabilidad entre calidades llegando en algunos casos a tasas de acierto del 100 %. Sin embargo estos valores no tienen en cuenta que la validación cruzada, método habitualmente utilizado para estimar la tasa de acierto global, devuelve un porcentaje de acierto optimista, con gran varianza (debido a muestras de datos pequeñas y modelos con covarianza alta entre ellos ya que solo difieren en una observación en su entrenamiento) y que en muchos casos la estructura de los datos (con una proporción elevada de datos en una calidad de estación) hace que la asignación de todas las observaciones a una misma calidad (la más frecuente) ya supere el 50 % de tasa de acierto global. Además, estas tasas de acierto están influidas por el número de calidades definidas en cada trabajo (siendo más alta la tasa de acierto global cuando se definen menos calidades de estación). Para corregir este sesgo optimista, se puede calcular el coeficiente de Kappa de Cohen (1960) que es la proporción de asignación correctas corregida por el número que se hubiese obtenido si hubiésemos asignado todas las observaciones a la calidad más frecuente. Este coeficiente puede tomar valores entre -1 y 1; si el valor es negativo indica que el método de clasificación es peor que el azar, 0 que es igual que el azar y 1 que la asignación es perfecta. En nuestro caso el coeficiente de la Kappa de Cohen (1960) es igual a 0,2309. Es decir que es el método seleccionado es un 23,09 % mejor que hacer la asignación a la calidad más frecuente. Se pueden obtener mejores resultados si se definen menos clases de calidades de estación y si las clases están representadas de forma equilibrada. Con este mismo conjunto de datos y también cuatro calidades, Bravo y Montero (2001) obtuvieron un coeficiente de Kappa de Cohen igual a 0,1624. Usando datos con una distribución más equilibrada de datos por clases diamétricas se obtienen valores de coeficiente de Kappa de Cohen (*k*) más alto como en el caso de Bravo et al (2011) que, en pinares de *Pinus pinea* en Calabria y 3 calidades de estación, obtuvieron un valor de *k* igual a 0,3778 mientras que Bravo-Oviedo y Montero (2005), también en masas de pino piñonero pero en Huelva y solo dos calidades, obtuvieron un valor de 0,4584. Finalmente, en rodales de pino silvestre en los páramos ácidos de Castilla y León, Bueis et al (2016) alcanzaron un valor de *k* igual a 0,5233. Es relevante señalar que este coeficiente (Cohen, 1960) es extremadamente conservador (muy duro) cuando la muestra es muy pequeña ya que asume que los valores obtenidos muestralmente representan fielmente la población. Cohen (1960) sugirió que la interpretación del coeficiente Kappa debiera ser el siguiente: $k \leq 0$ indica que no existe acuerdo, k entre 0,01 y 0,20 acuerdo nulo o bajo, entre 0,21 y 0,40 acuerdo débil, entre 0,41 y 0,60 acuerdo moderado, entre

0,61 y 0,80 acuerdo sustancial y entre 0,81 y 1 acuerdo casi perfecto. Como se puede deducir en los estudios forestales analizados (incluyendo el presente) el acuerdo puede clasificarse entre débil y moderado.

Los métodos ensayados están entre los más populares para la clasificación de observaciones en grupos. Cuando las clases están bien separadas respecto a las variables explicativas (no es nuestro caso) y el número de observaciones en cada clase y los predictores se distribuyen normalmente (como es en nuestro caso), el análisis discriminante lineal permite asignar a las observaciones de una forma más estable que la regresión logística multinomial (James et al, 2015) Una alternativa a estos métodos es el análisis discriminante cuadrático que funciona mejor cuando se dispone de pocas observaciones ya que no asume que la frontera de discriminación tenga una forma determinada (James et al, 2015)

Desde un punto de vista práctico la clasificación de los rodales por calidades puede hacerse mediante un gráfico (fig. 1) donde se representen las calidades o mediante el cálculo de la probabilidad de clasificación en cada una de las calidades a partir de los valores promedio y desviación típica de las variables explicativas (tabla 6) asumiendo su distribución normal.

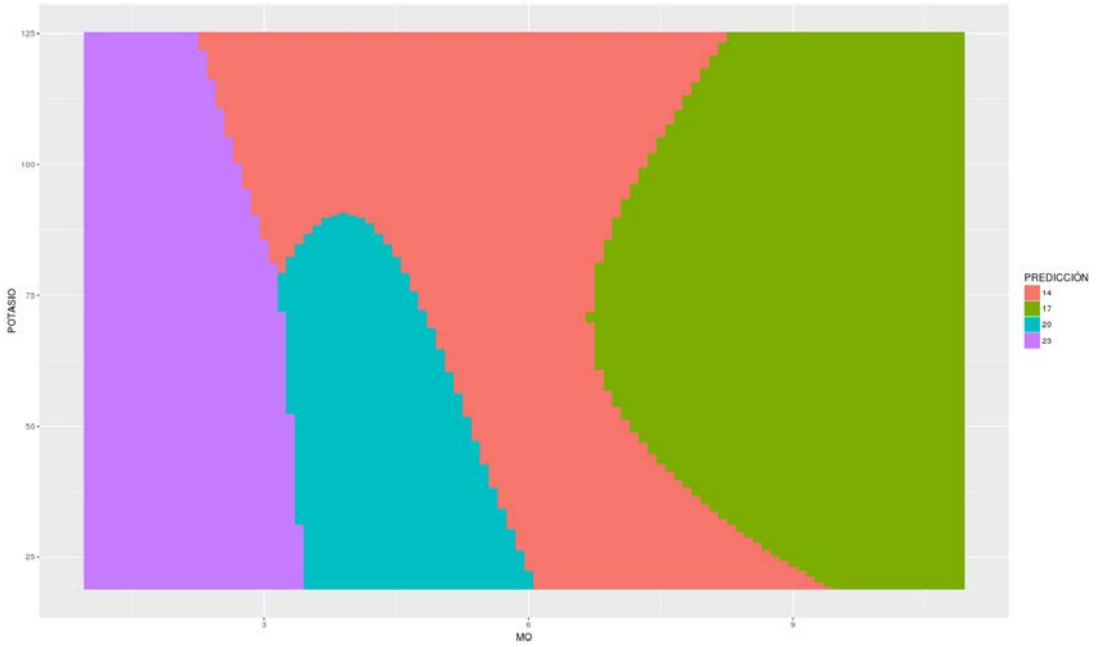

Figura 1. Clasificación por calidades (índice de sitio) usando la materia orgánica (%) y la concentración del potasio, no transformado, (en ppm) en rodales de pino silvestre en el Alto Ebro (norte de España)

Tabla 6 Promedio y desviación típica (entre paréntesis) para cada una de las calidades de estación de las variables que se incluyen en el modelo (5) desarrollado mediante Bayes ingenuo para clasificar la calidad de estación de rodales de Pinus sylvestris L. en el Alto Valle del Ebro.

Índice de sitio	√(K) ppm	Materia orgánica %
14	13,11 (3,54)	5,26 (1,59)
17	9,36 (1,60)	6,65 (2,88)
20	9,74 (4,20)	4,38 (0,95)
23	7,65 (3,52)	2,35 (1,79)

6. Conclusiones

Se han desarrollado reglas de clasificación de calidades de estación mediante tres métodos diferentes (Bayes ingenuo, análisis discriminante lineal y regresión logística multinomial) con datos de pinares de pino silvestre en Alto Valle del Ebro. El modelo estimado mediante Bayes ingenuo que usa como variables independientes el porcentaje de materia orgánica y la raíz cuadrada de la concentración del potasio (en ppm) es el más adecuado para clasificar los rodales por calidades de estación. La definición de la calidades, tanto en número como en límites de separación, y el equilibrio entre las muestras por cada una de ellas tiene impacto positivo (cuanto menor es el número de clases y mayor es el equilibrio entre clases) sobre la mejora de la precisión global (menor tasa de error) del modelo.

7. Bibliografía

BRAVO F, LUCÀ M, MERCURIO R, SIDARI M, MUSCOLO A, 2011. Soil and forest productivity: a case study from Stone pine (*Pinus pinea* L.) stands in Calabria (southern Italy). *iForest* 4: 25-30 disponible online en http://www.sisef.it/iforest/show.php?id=559

BRAVO, F., MONTERO, G., 2001. Site index estimation in Scots pine (*Pinus sylvestris* L.) stands in the High Ebro Basin (northern Spain) using soil attributes *Forestry* 74(4):395-406

BRAVO-OVIEDO, A., MONTERO, G. 2005. Site index in relation to edaphic variables in stone pine (*Pinus pinea* L.) stands in south west Spain. Ann. For. Sci. 62:61-72

BUEIS T, BRAVO F, PANDO V, TURRIÓN M-B. 2016. Relationship between environmental parameters and Pinus sylvestris L. site index in forest plantations in northern Spain acidic plateau. iForest 9: 394-401. – doi: 10.3832/ifor1600-008

COHEN, J.A. 1960. A coefficient of agreement for nominal scales. *Educational and Pshychological Measurements* 20:37-46

HAND, D.J., YU, K., 2001 Idiot's Bayes – Not so stupid after all? *International Statistical Review* 69(3):385-398

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. 2013 The elements of Statistical Learning. Data Mining, Inference and Prediction. Springer, 745 páginas. New York

IPCC 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. 151 páginas IPCC, Ginebra, Suiza (disponible online en https://www.ipcc.ch/report/ar5/syr/)

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. 2015 An introduction to Statistical Learning with applications in R. Springer 426 páginas. New York

JOKELA, E.J., WHITE, E.H., BERGLUND, J.V., 1988. Predicting Norway spruce growth from soil and topographic properties in New York. *Soil Science Society of America Journal* 52(3):809-815

MEYER, D., DIMITRIADOU, E., HORNIK, K., WEINGESSEL, A., LEISCH, F., CHANG, C-C 2015. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien R- package version 1.6-7 disponible en https://cran.r-project.org/web/packages/e1071/index.html

PAULO, J. A., FAIAS, S., GOMES, A. A., PALMA, J., TOMÉ, J., TOMÉ, M. 2015. Predicting site index from climate and soil variables for cork oak (Quercus suber L.) stands in Portugal. New Forests 46 (2): 293-307. DOI: http://dx.doi.org/10.1007/s11056-014-9462-4

R CORE TEAM; 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

SANCHEZ-RODRIGUEZ F, RODRIGUEZ-SOALLEIRO R, ESPANOL E, LOPEZ CA, MERINO A. 2002. Influence of edaphic factors and tree nutritive status on the productivity of Pinus radiata D. Don plantations in northwestern Spain. *Forest Ecol Manag* 171:181-189. doi:10.1016/s0378-1127(02)00471-1

SKOVSGAARD, J. P., VANCLAY, J.K. 2007. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. *Forestry* 81 (1): 13-31. doi: 10.1093/forestry/cpm041

SPIECKER, H, MIELIKÄINEN, K, KÖHL, M, SKOVSGAARD, JP. (Eds) 1996. Growth Trends of European Forests Springer Verlag, Berlin

VENABLES, W. N., RIPLEY, B. D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York