RESULTADOS

R03.- Tratamiento de datos para la clasificación de dientes de dinosaurio

Resumen	
TÍTULO:	R03.- Tratamiento de datos para la clasificación de dientes de dinosaurio (v. 2 .0)
RESUMEN:	El presente documento corresponde al resultado de la línea 2 del proyecto —relativa a la descripción e implementación de las técnicas de clasificación de los dientes—, en él se describen el listado de técnicas analizadas, ejemplos de procedimientos para el tratamiento de datos en ámbitos similares a nuestro caso de estudio y la implantación de procedimiento para el análisis de la información paleontológica que se ha desarrollado orientado al problema de la clasificación de los fósiles de dientes de dinosaurio. Los aspectos tratados en este protocolo son: a) Revisión de técnicas matemáticas para la clasificación de elementos. b) Ejemplos de laboratorio de datos en ámbitos afines. c) Diseño, descripción e implementación de procedimiento para el tratamiento de la base de datos con los dientes del yacimiento de Laño. d) Presentación de resultados. e) Análisis y discusión.
REDACCIÓN Y REVISIÓN:	María Álvarez Sáinz Iñaki López Ferrero Álvaro Rodríguez Miranda Angélica Torices Hernández Leire Usategui Frías
FECHA:	• Versión inicial (v.1.0) → noviembre de 2022 • Versión final (v. 2.0) → diciembre de 2022
REPOSITORIO:	http://hdl.handle.net/10810/58547

Universidad del País Vasco Euskal Herriko Unibertsitatea | Matematika Aplikatua Saila Departamento de Matemática Aplicada
Escuela de Ingeniería de Vitoria-Gasteiz
C/ Nieves Cano, 12. 01006, Vitoria-Gasteiz
Tfno: +34 945 01 3220
e-mail: alvaro.rodriguezm@ehu.eus

Universidad del País Vasco Euskal Herriko Unibertsitatea | Politika Publikoak eta Historia Ekonomikoa Saila Departamento de Políticas Públicas e Historia Económica
Facultad de Economía y Empresa
Avda. Lehendakari Agirre, 83. 48015, Bilbao
Tfno: +34 946 01 7097
e-mail: maria.alvarezsainz@ehu.eus

1.- Introducción

Comencemos con una descripción somera del análisis que sobre la colección de Laño se realiza en el artículo de referencia (Isasmendi *et al*., 2022)[1]. En este trabajo se parte de un esquema morfométrico que determina una serie de medidas sobre los dientes[2], conforme al siguiente esquema.

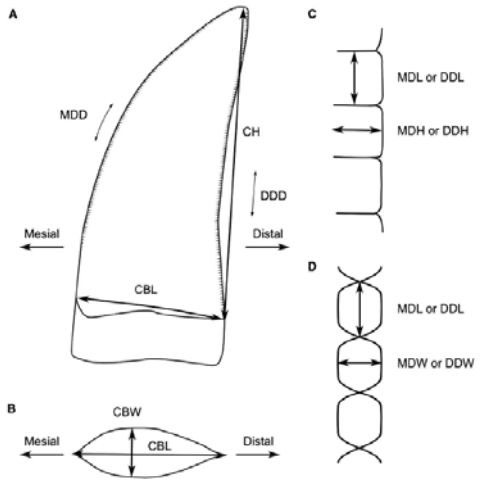


FIG. 2. Morphometric terminology used for the Laño teeth. A, theropod tooth in lateral or lingual view. B, theropod tooth in basal view. C, denticles in labial or lingual view. D, denticles in mesial or distal view. *Abbreviations*: CBL, crown base length; CBW, crown base width; CH, crown height; DDD, distal denticle density; DDH, distal denticle height; DDL, distal denticle length; DDW, distal denticle width; MDD, mesial denticle density; MDH, mesial denticle height; MDL, mesial denticle length; MDW, mesial denticle width.

Fig. 1.- Esquema de medias sobre los dientes. Corresponde a la figura número 2 del artículo (Isasmendi *et al*., 2022).

Como puede apreciarse, en todos los casos, se trata de dientes de animales carnívoros cuya morfología básica está adaptada a la funcionalidad de penetrar y rasgar la carne (forma de cuchillo). Los dinosaurios carnívoros reemplazaban los dientes a lo largo de toda su vida, por lo que los dientes aislados son uno de los fósiles más habituales que pueden encontrarse en el registro paleontológico (de ahí su interés para el estudio de las diferentes familias y especies). Una de las utilidades posibles es la de poder inferir la presencia de un tipo concreto de animal por la presencia de los fósiles de sus dientes; sin embargo, para poderlo hacer, es necesario identificarlos correctamente, para lo cual se deben establecer relaciones con ejemplares que sí

[1] Isasmendi, E., Torices, A., Canudo, J.I., Currie, P.J. and Pereda-Suberbiola, X. (2022), Upper Cretaceous European theropod palaeobiodiversity, palaeobiogeography and the intra-Maastrichtian faunal turnover: new contributions from the Iberian fossil site of Laño. *Papers in Palaeontology*, 8: e1419. https://doi.org/10.1002/spp2.1419

[2] Las referencias que se indican al respecto de las medidas elegidas son:

- Currie, P.J., Rigby, J.K. & Sloan, R.E. (1990) Theropod teeth from the Judith River Formation of southern Alberta, Canada. 107-125. In Carpenter, K & Currie, P.J. (eds.) *Dinosaur systematics: Perspectives and approaches*. Cambridge University Press, 356 pp.
- Smith, J.B., Vann, D.R. & Dodson, P. (2005) Dental morphology and variation in theropod dinosaurs: implications for the taxonomic identification of isolated teeth. *The Analytical Record*, 285A: 699-736.
- Hendricks, C. Matheus, O & Araújo, R. (2015) A proposed terminology of theropod teeth (Dinosauria, Saurischia). *Journal of Vertebrate Paleontology*, 35: e982797.

que se hayan encontrado en fósiles más completos en los que los dientes se hallasen en conexión anatómica con los cráneos (este tipo de fosilización de ejemplares completos es muy excepcional… y no se ha encontrado en el conjunto de Laño).

La base de datos adjunta al artículo reporta las medidas realizadas sobre 227 ejemplares de dientes, sobre los que se han considerado hasta 27 medidas distintas. No obstante, esta base de datos tiene muchos huecos debidos a que, por un lado, algunos especímenes no cuentan con algunas de las características consideradas (por ejemplo, en el caso de que los dientes sean lisos, no existen valores para la densidad de dentículos[3]); por otro lado, muchos de los ejemplares están incompletos, por lo que algunas de las medidas no pueden obtenerse, en algunos casos pueden estimarse a partir de la parte recuperada (lo cual debe indicarse en la base de datos con el fin de poder identificar que estos valores pueden no ser fiables) pero, en otros casos, serán valores ausentes.

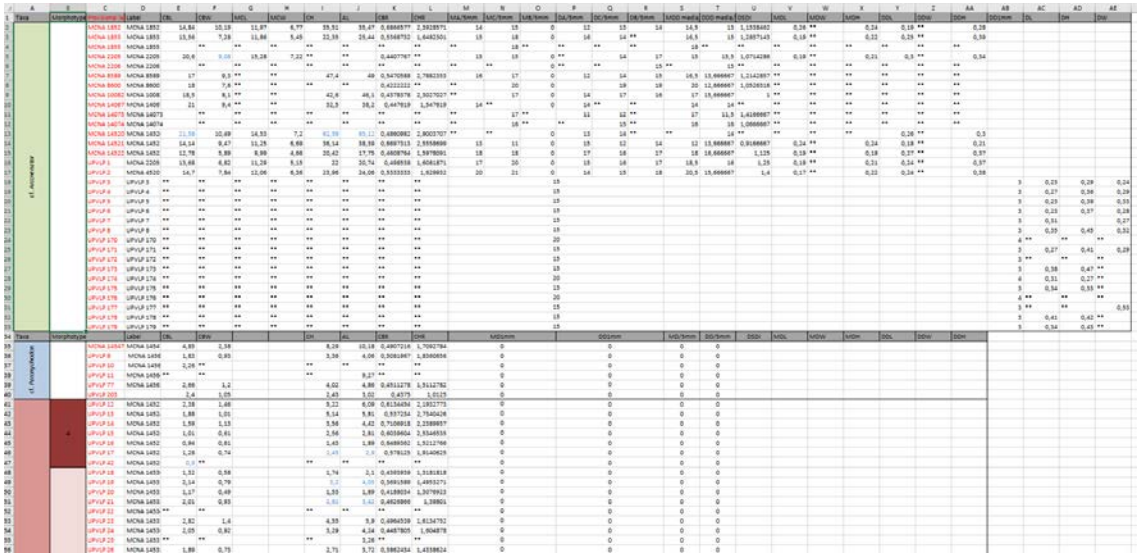


Fig. 2.- Vista parcial de la base de datos con medidas de los especímenes del yacimiento de Laño. Corresponde al material auxiliar del artículo (Isasmendi *et al*., 2022).

En esta base de datos, los especímenes ya están preclasificados. Esta asignación inicial corresponde a un «análisis experto» basado en el aspecto visual de cada pieza y en el que intervienen las dimensiones y la forma general (por ejemplo, la curvatura) del diente. Por otro lado, como se ha avanzado, la presencia y forma de los dentículos también juegan un papel esencial en la clasificación, así como otros factores como la rugosidad, estriaciones o estructura de las superficies[4]. No obstante, es importante notar que varias de estas características no

[3] Los «dentículos» son las aserraciones que presentan algunos dientes en sus filos. Se trata de una característica fundamental en los criterios de clasificación, tanto la presencia/ausencia de dentículos en los filos anteriores y posteriores como la forma de estos (si son triangulares, redondeados, etc.). Como puede verse, estos atributos son de naturaleza «cualitativa» y suelen soslayarse con el fin de poder aplicar algoritmos basados exclusivamente en caracteres «cuantitativos» lo que supone una pérdida importante de información relevante y, de hecho, es una de las preocupaciones principales que dieron origen al presente proyecto.

[4] A este respecto, véase, por ejemplo:

- Virág, A. & Ösi, A. (2017) Morphometry, microstucture, and wear pattern of neornithischian dinosaur teeth from the Upper Cretaceous Iharkút Locality (Hungary). *The Anatomical Record*, 300 (8): 1439-1463. https://doi.org/10.1002/ar.23592

aparecen medidas en la base de datos que se va a utilizar posteriormente para analizar la similaridad entre ejemplares y establecer así propuestas de clasificación automática.

En otro orden de ideas, ha de tenerse en cuenta que no existe una completa unanimidad en la asignación de muchos de los fósiles a una y otra especie[5] de dinosaurio. De hecho, el artículo que estamos utilizando reasigna varios especímenes a nuevas categorías respecto a estudios previos sobre los dientes de esta colección.

Siguiendo con este aspecto de la clasificación previa, es preciso indicar que las separaciones entre algunas de las categorías taxonómicas están sujetas a debate y que incluso existen algunas categorías que ni siquiera están reconocidas como válidas con carácter general. Por supuesto, conocer si un *Arcovenator* era un animal que debe aparecer muy claramente diferenciado de un espécimen de la familia de los *Dromaeosauriae* o si, por el contrario, se trata de dos subfamilias estrechamente relacionadas —o incluso si una denominación resulta ser una especialización de la otra— forma parte del conocimiento temático (paleontológico) que debe estar presente en todo el análisis y la interpretación de los datos[6]. Este hecho resulta relevante ya que una hipótesis común de muchas herramientas de clasificación es que las clases que se establecen corresponden a una partición exhaustiva y mutuamente excluyente del espacio de posibles soluciones (es decir, que todo elemento a clasificar será asignado a una única clase de salida) y, además, también suele ser razonable considerar que el nivel de especificación de todas las posibles clases sea similar… características éstas que, como se ha indicado, no siempre se cumplen en la asignación previa de los especímenes.

Para las fases posteriores de cálculo, del conjunto de dientes de la colección de Laño, se considerarán 66 ejemplares (distribuidos en 5 clases). Por otro lado, con el fin de establecer una comparativa conjunta, los atributos se reduce a cinco (que resultan comunes a todos ellos y que además van a estar presentes en los especímenes de comparación procedentes de otros yacimientos[7]).

[5] En el contexto explicativo de este documento, se hará una referencia genérica a las diferentes categorías de la base de datos como «especies» de dinosaurios. En realidad, esto no es exacto ya que dentro de la «taxonomía biológica» el concepto de «especie» tiene un significado concreto en lo relativo al grado de diferenciación que no siempre coincide con el nivel que corresponde los diferentes descriptores que se emplean (que, en algunos casos pueden estar referidos a una «familia», a un «género», etc.).

[6] Téngase en cuenta que dentro de las taxonomías existen diferentes niveles. En este problema, partimos del hecho de que todos los dientes pertenecen a terópodos, pero esta denominación está a nivel de «suborden» quedando aún por debajo niveles más específicos: infraorden, familia, subfamilia, género y especie. A la hora de presentar los especímenes, en algunos casos se afina más (es decir, se llega a identificar un nivel más detallado) y en otros aparecen asignaciones más genéricas (a niveles más altos) y también es posible encontrar en las bases de datos referencias a niveles diferentes para distintos especímenes fósiles.

[7] En este punto también existen discrepancias ya que los diversos autores pueden utilizar diferentes medidas para caracterizar los fósiles (o tener criterios diferentes para establecer una misma medida), por este motivo, las comparaciones entre tablas de medidas pueden tener un cierto riesgo de estar mezclando datos que no sean directamente comparables. Al respecto de los datos de referencia compilados de múltiples trabajos individuales, también es posible enfrentarse a un panorama heterogéneo y no exento de sesgos debidos a las clasificaciones publicadas por los diversos autores, así como condicionamientos cruzados (al basarse unos estudios en otros).

	A	B	C	D	E	F	G	H
1				Raw data				
2	Taxa	Reference	Specimen	CBL	CBW	CH	MDD	DDD
696	cf. Arcovenator sp. from Laño	This study	MCNA 1852	14,8	10,2	35,51	15	13
697	cf. Arcovenator sp. from Laño	This study	MCNA 1853	13,6	7,28	22,35	18	14
698	cf. Arcovenator sp. from Laño	This study	MCNA 8589	17	9,3	47,4	17	14
699	cf. Arcovenator sp. from Laño	This study	MCNA 10082	18,5	8,1	42,6	17,00	17,00
700	cf. Arcovenator sp. from Laño	This study	MCNA 14521	14,1	9,47	36,14	11	12
701	cf. Arcovenator sp. from Laño	This study	MCNA 14522	12,8	5,89	20,42	18,00	16
702	cf. Arcovenator sp. from Laño	This study	MCNA 22051	13,7	6,82	22	20	16
703	cf. Arcovenator sp. from Laño	This study	MCNA 4520	14,7	7,84	23,96	21,00	15
704	cf. Paronychodon sp. from Laño	This study	MCNA 14547	4,85	2,38	8,29	0	0
705	cf. Paronychodon sp. from Laño	This study	MCNA 14562	1,83	0,93	3,36	0	0
706	cf. Paronychodon sp. from Laño	This study	UPVLP 77	2,66	1,2	4,02	0	0
707	cf. Paronychodon sp. from Laño	This study	UPVLP 203	2,4	1,05	2,43	0	0
708	Paraves indet. from Laño	This study	MCNA 14523	2,38	1,46	5,22	0	0
709	Paraves indet. from Laño	This study	MCNA 14524	1,88	1,01	5,14	0	0
710	Paraves indet. from Laño	This study	MCNA 14525	1,59	1,13	3,56	0	0
711	Paraves indet. from Laño	This study	MCNA 14526	1,01	0,61	2,56	0	0
712	Paraves indet. from Laño	This study	MCNA 14527	0,94	0,61	1,43	0	0
713	Paraves indet. from Laño	This study	MCNA 14528	1,28	0,74	2,45	0	0
714	Paraves indet. from Laño	This study	MCNA 14530	1,32	0,58	1,74	0	0
715	Paraves indet. from Laño	This study	MCNA 14531	2,14	0,79	3,2	0	0
716	Paraves indet. from Laño	This study	MCNA 14532	1,17	0,49	1,53	0	0
717	Paraves indet. from Laño	This study	MCNA 14533	2,01	0,93	2,81	0	0
718	Paraves indet. from Laño	This study	MCNA 14535	2,82	1,4	4,55	0	0
719	Paraves indet. from Laño	This study	MCNA 14536	2,05	0,92	3,29	0	0
720	Paraves indet. from Laño	This study	MCNA 14538	1,89	0,73	2,71	0	0
721	Paraves indet. from Laño	This study	MCNA 14539	1,42	0,76	2,52	0	0
722	Paraves indet. from Laño	This study	MCNA 14540	1,36	0,61	1,86	0	0
723	Paraves indet. from Laño	This study	MCNA 14541	1,67	0,55	1,87	0	0
724	Paraves indet. from Laño	This study	MCNA 14542	1,29	0,56	1,53	0	0
725	Paraves indet. from Laño	This study	MCNA 14544	1,01	0,56	1,26	0	0
726	Paraves indet. from Laño	This study	MCNA 14548	1,17	0,82	1,36	0	0
727	Paraves indet. from Laño	This study	MCNA 14550	1,11	0,6	1,58	0	0
728	Paraves indet. from Laño	This study	MCNA 14552	0,93	0,5	1,3	0	0
729	Paraves indet. from Laño	This study	MCNA 14557	1,14	0,76	2,15	0	0
730	Paraves indet. from Laño	This study	MCNA 14558	1,76	0,8	2,94	0	0
731	Paraves indet. from Laño	This study	MCNA 14559	1,3	0,59	1,98	0	0
732	cf. Dromaeosaurinae indet. from Laño	This study	MCNA 14623	4,5	2,13	5,66	40	30
733	cf. Dromaeosaurinae indet. from Laño	This study	MCNA 14624	5,29	4,32	6,62	45	30
734	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14606	3,18	1,22	4,33	40	35
735	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14608	2,96	1,03	3,28	0	35
736	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14610	1,34	0,7	1,92	50	45
737	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14619	1,94	0,6	3,78	0	50
738	cf. *Richardoestesia* sp. from Laño	This study	MCNA 1993-184 16	2,99	1,51	6,82	0	35
739	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14566	2,78	1,21	5,02	0	40
740	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14568	1,72	0,69	2,43	55	40
741	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14570	1,71	0,6	2,3	0	75
742	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14571	1,8	0,62	1,89	0	80
743	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14572	1,46	0,55	2,53	0	45
744	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14573	1,69	0,85	3,03	65	55
745	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14580	1,53	0,56	2,21	0	75

Fig. 3.- Conjunto de 66 dientes del yacimiento de Laño seleccionados para los análisis posteriores, nótense algunos valores estimados que aparecen representados en color azul, por ejemplo, el valor CH de las filas 713, 715, 717, 721… (Isasmendi *et al*., 2022).

De estos cinco parámetros cuantitativos, los tres primeros (CBL, CBW y CH) corresponden a las medidas volumétricas principales (similares a lo que sería el «largo», «ancho» y «alto» pero teniendo en cuenta las zonas del diente en que se miden estos valores) que se presentan en milímetros. Por otro lado, los dos parámetros restantes (MDD y DDD) son las densidades de dentículos en los filos anterior y posterior[8] (en esta tabla, se recoge el número de dentículos cada 5 mm).

Al tratarse de datos de diferente naturaleza (distancias los tres primeros y densidades los dos últimos) que se quieren combinar, es necesario transformarlos a una versión «normalizada». El método seleccionado —sobre el que discutiremos más adelante— es la toma de logaritmos; pero —como se da el caso de que los valores que figuran en las densidades de los especímenes

[8] Aquí existe un nuevo punto de conflicto ya que las densidades no siempre son constantes a lo largo del filo. Algunos autores establecen varias medidas, separando las densidades en la parte alta, media y baja del diente… pero otros utilizan un valor promedio.

sin dentículos (bordes lisos) son ceros[9] y que el logaritmo de cero no es un valor finito— se aplica una modificación que consiste en añadir una unidad al valor original antes de tomar el logaritmo[10], es decir:

valor normalizado = logaritmo (valor original + 1)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1				Raw data					Normalized data				
2	Taxa	Reference	Specimen	CBL	CBW	CH	MDD	DDD	CBL	CBW	CH	MDD	DDD
696	cf. Arcovenator sp. from Laño	This study	MCNA 1852	14,8	10,2	35,51	15	13	1,19975518	1,04883009	1,56241183	1,20411998	1,14612804
697	cf. Arcovenator sp. from Laño	This study	MCNA 1853	13,6	7,28	22,35	18	14	1,16316137	0,91803034	1,36828688	1,2787536	1,17609126
698	cf. Arcovenator sp. from Laño	This study	MCNA 8589	17	9,3	47,4	17	14	1,25527251	1,01283722	1,68484536	1,25527251	1,17609126
699	cf. Arcovenator sp. from Laño	This study	MCNA 10082	18,5	8,1	42,6	17,00	17,00	1,29003461	0,95904139	1,63948649	1,25527251	1,25527251
700	cf. Arcovenator sp. from Laño	This study	MCNA 14521	14,1	9,47	36,14	11	12	1,18012588	1,01994668	1,5698419	1,07918125	1,11394335
701	cf. Arcovenator sp. from Laño	This study	MCNA 14522	12,8	5,89	20,42	18,00	16	1,13924922	0,83821922	1,33081947	1,2787536	1,23044892
702	cf. Arcovenator sp. from Laño	This study	MCNA 22051	13,7	6,82	22	20	16	1,16672606	0,89320675	1,36172784	1,32221929	1,23044892
703	cf. Arcovenator sp. from Laño	This study	MCNA 4520	14,7	7,84	23,96	21,00	15	1,19589965	0,94645227	1,39724458	1,34242268	1,20411998
704	cf. Paronychodon sp. from Laño	This study	MCNA 14547	4,85	2,38	8,29	0	0	0,76715587	0,5289167	0,96801571	0	0
705	cf. Paronychodon sp. from Laño	This study	MCNA 14562	1,83	0,93	3,36	0	0	0,45178644	0,28555731	0,63948649	0	0
706	cf. Paronychodon sp. from Laño	This study	UPVLP 77	2,66	1,2	4,02	0	0	0,56348109	0,34242268	0,70070372	0	0
707	cf. Paronychodon sp. from Laño	This study	UPVLP 203	2,4	1,05	2,43	0	0	0,53147892	0,31175386	0,53529412	0	0
708	Paraves indet. from Laño	This study	MCNA 14523	2,38	1,46	5,22	0	0	0,5289167	0,39093511	0,79379038	0	0
709	Paraves indet. from Laño	This study	MCNA 14524	1,88	1,01	5,14	0	0	0,45939249	0,30319606	0,78816837	0	0
710	Paraves indet. from Laño	This study	MCNA 14525	1,59	1,13	3,56	0	0	0,41329976	0,3283796	0,65896484	0	0
711	Paraves indet. from Laño	This study	MCNA 14526	1,01	0,61	2,56	0	0	0,30319606	0,20682588	0,55145	0	0
712	Paraves indet. from Laño	This study	MCNA 14527	0,94	0,61	1,43	0	0	0,28780173	0,20682588	0,38560627	0	0
713	Paraves indet. from Laño	This study	MCNA 14528	1,28	0,74	2,45	0	0	0,35793485	0,24054925	0,5378191	0	0
714	Paraves indet. from Laño	This study	MCNA 14530	1,32	0,58	1,74	0	0	0,36548798	0,19865709	0,43775056	0	0
715	Paraves indet. from Laño	This study	MCNA 14531	2,14	0,79	3,2	0	0	0,49692965	0,25285303	0,62324929	0	0
716	Paraves indet. from Laño	This study	MCNA 14532	1,17	0,49	1,53	0	0	0,33645973	0,17318627	0,40312052	0	0
717	Paraves indet. from Laño	This study	MCNA 14533	2,01	0,93	2,81	0	0	0,4785665	0,28555731	0,58092498	0	0
718	Paraves indet. from Laño	This study	MCNA 14535	2,82	1,4	4,55	0	0	0,58206336	0,38021124	0,74429298	0	0
719	Paraves indet. from Laño	This study	MCNA 14536	2,05	0,92	3,29	0	0	0,48429984	0,28330123	0,63245729	0	0
720	Paraves indet. from Laño	This study	MCNA 14538	1,89	0,73	2,71	0	0	0,46089784	0,2380461	0,56937391	0	0
721	Paraves indet. from Laño	This study	MCNA 14539	1,42	0,76	2,52	0	0	0,38381537	0,24551267	0,54654266	0	0
722	Paraves indet. from Laño	This study	MCNA 14540	1,36	0,61	1,86	0	0	0,372912	0,20682588	0,45636603	0	0
723	Paraves indet. from Laño	This study	MCNA 14541	1,67	0,55	1,87	0	0	0,42651126	0,1903317	0,4578819	0	0
724	Paraves indet. from Laño	This study	MCNA 14542	1,29	0,56	1,53	0	0	0,35983548	0,1931246	0,40312052	0	0
725	Paraves indet. from Laño	This study	MCNA 14544	1,01	0,56	1,26	0	0	0,30319606	0,1931246	0,35410844	0	0
726	Paraves indet. from Laño	This study	MCNA 14548	1,17	0,82	1,36	0	0	0,33645973	0,26007139	0,372912	0	0
727	Paraves indet. from Laño	This study	MCNA 14550	1,11	0,6	1,58	0	0	0,32428246	0,20411998	0,41161971	0	0
728	Paraves indet. from Laño	This study	MCNA 14552	0,93	0,5	1,3	0	0	0,28555731	0,17609126	0,36172784	0	0
729	Paraves indet. from Laño	This study	MCNA 14557	1,14	0,76	2,15	0	0	0,33041377	0,24551267	0,49831055	0	0
730	Paraves indet. from Laño	This study	MCNA 14558	1,76	0,8	2,94	0	0	0,44090908	0,25527251	0,59549622	0	0
731	Paraves indet. from Laño	This study	MCNA 14559	1,3	0,59	1,98	0	0	0,36172784	0,20139712	0,47421626	0	0
732	cf. Dromaeosaurinae indet. from Laño	This study	MCNA 14623	4,5	2,13	5,66	40	30	0,74036269	0,49554434	0,82347423	1,61278386	1,49136169
733	cf. Dromaeosaurinae indet. from Laño	This study	MCNA 14624	5,29	4,32	6,62	45	30	0,79865065	0,72591163	0,88195497	1,66275783	1,49136169
734	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14606	3,18	1,22	4,33	40	35	0,62117628	0,34635297	0,72672721	1,61278386	1,5563025
735	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14608	2,96	1,03	3,28	0	35	0,59769519	0,30749604	0,63144377	0	1,5563025
736	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14610	1,34	0,7	1,92	50	45	0,36921586	0,23044892	0,46538285	1,70757018	1,66275783
737	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14619	1,94	0,6	3,78	0	50	0,46834733	0,20411998	0,6794279	0	1,70757018
738	cf. *Richardoestesia* sp. from Laño	This study	MCNA 1993-184 16	2,99	1,51	6,82	0	35	0,6009729	0,39967372	0,89320675	0	1,5563025
739	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14566	2,78	1,21	5,02	0	40	0,5774918	0,34439227	0,77959649	0	1,61278386
740	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14568	1,72	0,69	2,43	55	40	0,4345689	0,2278867	0,53529412	1,74818803	1,61278386
741	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14570	1,71	0,6	2,3	0	75	0,43296929	0,20411998	0,51851394	0	1,88081359
742	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14571	1,8	0,62	1,89	0	80	0,44715803	0,20951501	0,46089784	0	1,90848502
743	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14572	1,46	0,55	2,53	0	45	0,39093511	0,1903317	0,54777471	0	1,66275783
744	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14573	1,69	0,85	3,03	65	55	0,42975228	0,26717173	0,60530505	1,81954394	1,74818803
745	cf. *Richardoestesia* sp. from Laño	This study	MCNA 14580	1,53	0,56	2,21	0	75	0,40312052	0,1931246	0,50650503	0	1,88081359

Fig. 4.- Vista parcial de la base de datos con medidas de los especímenes del yacimiento de Laño. Corresponde al material auxiliar del artículo (Isasmendi *et al*., 2022).

Además de los datos de los dientes procedentes del yacimiento de Laño, la base de datos utilizada en el estudio incluye la información otros 692 dientes de referencia tomados de diferentes trabajos y que presentan ejemplares de 16 clases —tipos de dinosaurios— (que incluyen las 5 utilizadas en Laño).

[9] Este es otro aspecto que merecerá una reflexión más profunda a lo largo del texto. Considerar que los dientes lisos tienen una densidad de cero dentículos como una posibilidad más dentro de una sucesión de posibles valores: 0, 1, 2, 3… tiene implicaciones diferentes a considerar que el hecho de ser «liso»/«aserrado» es una variable dicotómica (de tipo «dentículos No/Sí, {0, 1}...»).

[10] Como el logaritmo de uno es cero, los valores que antes de normalizar tenían el valor de cero, se transforman en: logaritmo(1) = 0. Es decir, que mantienen su valor igual a cero.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1				Raw data					Normalized data				
2	Taxa	Reference	Specimen	CBL	CBW	CH	MDD	DDD	CBL	CBW	CH	MDD	DDD
3	*Pyroraptor olympius*	RonanAllain pers.	?	6,1	3,1	16	42,5	30	0,85125835	0,61278386	1,23044892	1,63848926	1,49136169
4	cf. *Arcovenator* Armuña	Pérez-García et al., 2016	UPUAM 14044	21,3	11,4	45,5	15	15	1,34830486	1,09342169	1,66745295	1,20411998	1,20411998
5	cf. *Arcovenator* Armuña	Pérez-García et al., 2016	UPUAM 14047	21	9,2	41	15	15	1,34242268	1,00860017	1,62324929	1,20411998	1,20411998
6	cf. *Arcovenator* Armuña	Pérez-García et al., 2016	UPUAM 14048	14	6,4	19,1	15	15	1,17609126	0,86923172	1,30319606	1,20411998	1,20411998
7	*Arcovenator escotae*	Tortosa et al., 2014	MHNAIX-PV-2011-12-15	15	8	52,65	17,5	20	1,20411998	0,95424251	1,72956973	1,26717173	1,32221929
8	*Arcovenator escotae*	Tortosa et al., 2014	MHNAIX-PV-2011-12-187	19	10	51	17,5	17,5	1,30103	1,04139269	1,71600334	1,26717173	1,26717173
9	*Majungasaurus*	Smith et al., 2005	FMNHPR2008	13	9,46	30,11	11	11,5	1,14581771	1,01953168	1,49290001	1,07918125	1,09691001
10	*Majungasaurus*	Smith et al., 2005	UA 8716	12,4	9,26	27,05	10	10	1,12742878	1,01114736	1,44793287	1,04139269	1,04139269
11	*Majungasaurus*	Smith et al., 2005	UA 8716	12,5	8,3	27,69	10	11	1,13065535	0,96848295	1,45773055	1,04139269	1,07918125
12	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,3	8,62	36,9	8,5	9	1,28555731	0,98317507	1,57863921	0,97772361	1
13	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,4	9,21	38,08	9,5	9	1,28712962	1,00902574	1,59195456	1,0211893	1
14	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,9	8,86	35,54	10	10	1,2995073	0,99387691	1,56276854	1,04139269	1,04139269
15	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,2	9,1	38,68	9,7	9	1,28262211	1,00432137	1,59857166	1,02938378	1
16	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	7,88	3,47	12,45	12	14	0,94841297	0,65030752	1,12872228	1,11394335	1,17609126
17	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	8,81	7,2	19,88	9	9	0,99166901	0,91381385	1,31973049	1	1
18	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	13,3	8,56	25,37	11	9,5	1,15533604	0,98045789	1,42111013	1,07918125	1,0211893
19	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	14,2	7,72	25,13	11	11	1,18298497	0,94051648	1,41713941	1,07918125	1,07918125
20	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,7	7,12	19,93	12	10	1,1354507	0,90955603	1,32076923	1,11394335	1,04139269
21	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,5	6,69	19,21	11,7	10,5	1,1312978	0,88592634	1,30556631	1,10380372	1,06069784
22	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,3	6,7	17,87	12	11,3	1,12319808	0,88649073	1,2757719	1,11394335	1,08990511
23	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	11,7	6,29	16,19	12,8	10,5	1,10448711	0,86272753	1,23527588	1,13987909	1,06069784
24	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	9,33	5,36	14,48	15	11	1,01410032	0,80345712	1,18977096	1,20411998	1,07918125
25	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	10,9	8,48	22,88	8	8,5	1,07591176	0,97680834	1,37803432	0,95424251	0,97772361
26	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	13,5	8,27	22,87	9,5	9	1,16226561	0,96707973	1,37785242	1,0211893	1
27	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	13,9	7,9	24,08	9,5	10	1,17318627	0,94939001	1,39932753	1,0211893	1,04139269
28	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	13,5	7,77	25,75	9,5	10,5	1,161368	0,94299959	1,42732379	1,0211893	1,06069784
29	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,8	8,59	23,76	10,8	10	1,14050804	0,98181861	1,39375064	1,07188201	1,04139269
30	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	13,2	7,05	23	11	9,5	1,15136985	0,90579588	1,38021124	1,07918125	1,0211893
31	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	11,3	5,79	18,73	11,5	11,5	1,08849047	0,83186977	1,29512709	1,09691001	1,09691001
32	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,9	5,25	18,71	12,2	11,7	1,14238947	0,79588002	1,29468662	1,12057393	1,10380372
33	*Majungasaurus*	Smith et al., 2005	FMNH PR	16,9	8,67	37,63	10	9,3	1,25285303	0,98542647	1,58692471	1,04139269	1,01283722
34	*Majungasaurus*	Smith et al., 2005	FMNH PR	17,1	8,81	35,01	12	10	1,25767857	0,99166901	1,55642312	1,11394335	1,04139269
35	*Dromaeosaurus*	Hendrickx et al., 2015	AMNH 5356	7,95	4,84	17,97	17,5	18,8	0,95182304	0,76641285	1,27806733	1,26717173	1,2955671
36	*Dromaeosaurus*	Hendrickx et al., 2015	AMNH 5356	6,91	4,18	12,34	13,8	16,7	0,89817648	0,71432976	1,12515583	1,16879202	1,2469907
37	*Dromaeosaurus*	Hendrickx et al., 2015	AMNH 5356	5,99	4,45	12,8	17,5	20	0,84447718	0,7363965	1,13987909	1,26717173	1,32221929
38	*Dromaeosaurus*	Hendrickx et al., 2015	AMNH 5356	6,19	3,97	13,78	16,7	15,8	0,85672889	0,69635639	1,16967443	1,2469907	1,22608412
39	*Dromaeosaurus*	Hendrickx et al., 2015	AMNH 5356	5,57	3,26	10,14	15	20	0,81756537	0,6294096	1,04688519	1,20411998	1,32221929

Fig. 5.- Vista parcial de la base de datos con medidas de los especímenes de los yacimientos de referencia utilizados para la comparación. Corresponde al material auxiliar del artículo (Isasmendi *et al.*, 2022).

Con esta base de datos se realizan dos tipos de análisis: un «análisis de componentes principales» y un «análisis discriminante».

a) Análisis de componentes principales. Se extraen los dos primeros componentes (los cuales explican, respectivamente, el 64 % y el 23 % de la varianza, entre ambos un 87 %) con el fin de hacer una representación bidimensional del conjunto de elementos. También se determina que la primera componente principal está definida principalmente por las tres primeras variables (CH, CBL y CBW) —es decir por las que representan las dimensiones del diente— mientras que la segunda está especialmente definida por las variables relativas a la densidad de dentículos (especialmente DDD).

   El gráfico de dispersión resultante se presenta a continuación. No entramos ahora a comentarlo más allá de lo que se dice en el propio artículo, donde se identifican tres grandes grupos: el de los dientes con dientes lisos (abajo a la izquierda en el gráfico), los de dientes serrados pequeños (zona central y superior izquierda) y los de dientes serrados grandes (parte derecha del gráfico).

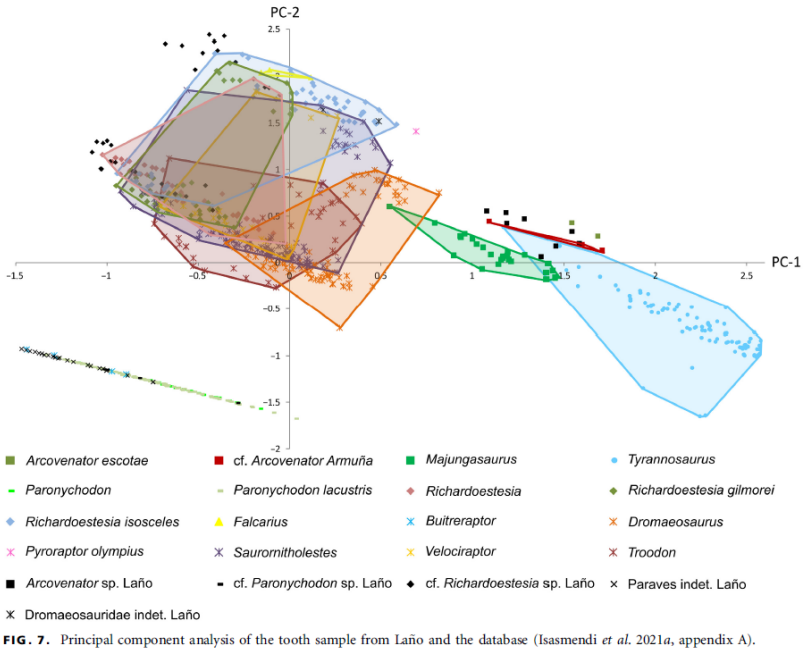


FIG. 7. Principal component analysis of the tooth sample from Laño and the database (Isasmendi *et al.* 2021*a*, appendix A).

Fig. 6.- Gráfico de dispersión que corresponde a los dos ejes factoriales del «análisis de componentes principales». Corresponde a la figura número 7 del artículo (Isasmendi *et al*., 2022).

b) <u>Análisis discriminante</u>. De este análisis se indica que el resultado resulta bastante bueno ya que el 85'5 % de los elementos aparecen clasificados en sus grupos de origen. En el artículo se indican cuáles son los errores de clasificación; no obstante, para alguien ajeno a la paleontología, resulta difícil indicar si son confusiones aceptables o errores de calado.

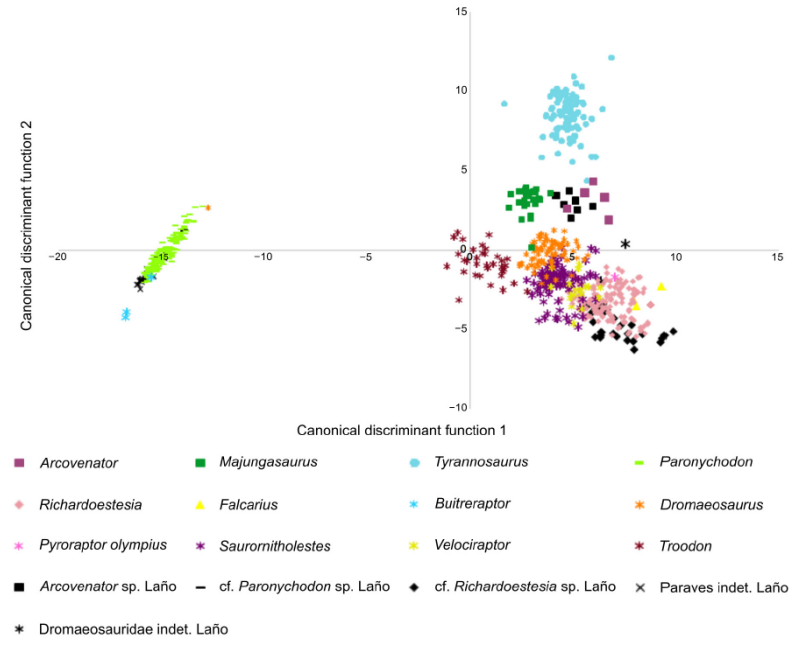


FIG. 8. Discriminant analysis of the tooth sample from Laño and the database (Isasmendi *et al.* 2021*a*, appendix A).

Fig. 7.- Gráfico de dispersión que corresponde a los dos ejes factoriales del «análisis discriminante». Corresponde a la figura número 8 del artículo (Isasmendi *et al*., 2022).

Respecto a la utilidad del estudio estadístico representado por los gráficos anteriores, se puede decir que se incorpora como argumento para reforzar la asignación previa realizada mediante criterio experto y para intentar intuir un posible método de diferenciación entre diferentes taxones.

Antes de finalizar esta introducción es oportuno indicar que los autores del artículo (y, en general toda la comunidad científica que aborda este campo) acepta el hecho de que existe cierta variabilidad en la forma y dimensiones de los dientes dentro de la misma especie —e incluso dentro de un mismo individuo— que se acentúa más si se plantea la existencia de cambios entre ejemplares jóvenes y adultos. Por otro lado, también se apunta la existencia de «dientes anómalos» como posible justificación de asignaciones de especímenes con diferencias significativas a otros elementos de la misma categoría.

2.- Objetivos

El trabajo descrito anteriormente constituye un ejemplo del uso validado de las dos técnicas estadísticas comentadas (análisis de componentes principales y análisis discriminante) en el ámbito de la paleontología. No obstante, a pesar de aportar luz sobre algunos aspectos, los propios paleontólogos ven estas herramientas con cierta insatisfacción ya que, por un lado, son plenamente conscientes que están dejando fuera del análisis cuantitativo una gran cantidad de información sobre los elementos analizados (los dientes en este caso) que resulta relevante en la clasificación experta. En segundo lugar, si bien los resultados muestran separaciones claras entre algunos grandes grupos de elementos, también presentan conjuntos con grandes solapes para los que, aparentemente, las técnicas empleadas no dan opciones de discriminación. En tercer lugar, estas herramientas tienen una fuerte componente de «caja negra» en la que no queda muy claro qué es lo que realmente ha sucedido durante el proceso de cálculo. Por último, los paleontólogos son plenamente conscientes de que los resultados están fuertemente condicionados por los datos de partida que han introducido, así que los aceptan con cierto recelo y temor de incurrir en el autoengaño, al fin y al cabo, conocen las limitaciones existentes en los datos de partida.

Por estos motivos, la línea de análisis matemático del presente proyecto tiene por objetivo adquirir el conocimiento suficiente para dar un asesoramiento a los paleontólogos que permita establecer un procedimiento de tratamiento de los datos cuya adecuación y fiabilidad pueda ser contrastada. Como resultado tangible, se ofrecerá un procedimiento de tratamiento de la información (parte analítica) que deberá ser integrado con diferentes fases de discusión y análisis (desde el conocimiento paleontológico).

Para ello, se va a partir de una revisión de diversas técnicas matemáticas útiles para la clasificación de elementos en una base de datos a partir de los valores recogidos de una serie de características. Se prestará también atención a técnicas que permitan trabajar con variables de tipo categórico (nominal) ya que son menos habituales dentro de los análisis actualmente realizados, a pesar de que el hecho de descartar esta información supone una merma en la capacidad de clasificación de los algoritmos actuales.

Estas técnicas deberán emplearse de forma contextualizada y, posiblemente, concatenadas de manera que puedan combinarse para obtener mejores resultados que cada una de manera individual. Como referencia de la combinación de herramientas, se van a presentar algunos ejemplos completos de procesamiento de datos (que abarcan desde el registro, pasando por el

tratamiento de la información a la generación y análisis de resultados) en otros ámbitos que pueden resultar ilustrativos, en concreto en los ámbitos de: la sedimentología y la arqueometría. Este análisis comparado permitirá ofrecer una perspectiva más amplia respecto a la forma de aplicar las técnicas de procesamientos de los datos, así como las posibilidades de análisis de la información y de los resultados.

Dado que el uso del procedimiento de procesado de la información paleontológica se utilizará en un contexto multidisciplinar, se considera necesario que dicho proceso sea sencillo de comunicar e interpretar. Por este motivo, se ha determinado exponerlo utilizando un criterio gráfico, a modo de diagrama de flujo en el cual las diferentes técnicas se consideran piezas recolocables en diferentes partes del esquema a modo de puzle.

Todo lo anterior debe considerarse en el marco temporal y los recursos disponibles en el presente proyecto que, evidentemente, no van a permitir agotar las posibilidades de investigación y desarrollo. Por lo tanto, se considera conveniente que este proyecto sirva como punto de partida a una posible extensión en empeños futuros.

3.- Esquema general

El desarrollo de la presente línea combina diferentes componentes:

a) Por un lado, se deben compilar las bases de datos de información paleontológica que puedan utilizarse para probar los algoritmos y comparar los resultados con los ya obtenidos en estudios previos. A este respecto se dispone de los datos correspondientes al artículo del yacimiento de Laño previamente indicado (trabajo realizado por el paleontólogo Xabier Pereda-Suberbiola) y otra base de datos adicional de otro yacimiento (Lo Hueco, en Cuenca) proporcionada por la paleontóloga Angélica Torices.

b) Resulta necesario hacer una revisión de la base matemática de las técnicas empleadas con el fin de ver cuál es su fundamento, condiciones de partida que deben cumplir los datos antes de ser procesados y la manera correcta de interpretar los resultados; así como la discusión sobre su aplicabilidad a los diferentes casos. A este respecto, se han seleccionado cuatro técnicas que cubren, a grandes rasgos, los principales ámbitos que es oportuno tratar, éstas son:

    - T01.- Análisis de componentes principales. Como metodología habitual dedicada a la reducción de la dimensionalidad del conjunto de datos, es decir, conseguir pasar de un elevado número de variables (columnas de atributos) a una versión más reducida de atributos (generados por combinación lineal de los primeros) que mantengan la mayor parte posible de la información (sobre la variabilidad entre los elementos). En particular, esta transformación facilita la generación de representaciones gráficas generando diagramas de dispersión entre las 2 primeras componentes.

    - T02.- Análisis de correspondencias. Tal como se ha indicado, una de las principales carencias detectadas en los análisis realizados es la falta de consideración de las variables de tipo cualitativo (categóricas). El análisis de correspondencias permite abordarlas mediante tablas de contingencia bidimensionales (a partir de las cuales se calculan perfiles que, posteriormente, pueden representarse gráficamente tras haber pasado por una reducción de

dimensionalidad mediante el análisis de componentes principales descrito en la técnica anterior).

- T03.- Análisis factorial discriminante. En este caso, los datos parten de una asignación previa a una clase y la variabilidad total del conjunto de datos se considera como la suma de dos componentes: una interna dentro de los propios grupos (entre los elementos preliminarmente identificados dentro de la misma clase) y otra entre los diferentes grupos. El cálculo trata de mostrar la situación (transformación a nuevos ejes factoriales) que mejor muestre la separación entre grupos. Este tipo de análisis permite validar las asignaciones a las categorías que se hayan establecido de manera previa, así como clasificar nuevos elementos a alguna de las categorías ya establecidas.

- T04.- Análisis de conglomerados (clúster). El análisis de conglomerados agrupa los elementos por su proximidad[11] sin considerar una clasificación previa definida de manera externa. Esto permite generar grupos desde cero, así como un contraste posterior con una clasificación experta que permita determinar su concordancia.

De cada una de estas técnicas se ha realizado una pequeña ficha basada en la revisión bibliográfica que describe brevemente su objeto y la forma de cálculo. Además, se han desarrollado ejemplos de cálculo con los datos de las bases de datos paleontológicas disponibles, de manera que se ha podido realizar una primera valoración sobre la posible utilidad de cada técnica.

Por supuesto, la lista anterior sería extensible en el futuro, mediante la incorporación de nuevas técnicas que se podrían ir estudiando. También resulta ampliable mediante profundización en cada una de las técnicas, ya que las fichas que se han realizado analizan los aspectos fundamentales, pero se debe considerar que cada una de estas técnicas cuenta también con aspectos avanzados que permiten sacar un mayor rendimiento a los datos y realizar una interpretación más afinada de los resultados.

c) Para disponer de una mayor perspectiva que nos permita saber cómo pueden utilizarse estas herramientas en el contexto de un procesamiento de datos desde el diseño de la captura de la información, el registro, preprocesado de los datos, análisis cuantitativo y cualitativo, hasta la obtención y discusión de los resultados resulta interesante ver otros ejemplos de aplicación en otras áreas. En concreto se han considerado dos casos: la clasificación de sedimentos costeros (con el fin de determinar las dinámicas de las playas) y la clasificación de piezas cerámicas arqueológica a través de caracterización química (con el objeto de establecer relaciones e identificar la proveniencia de las producciones cerámicas).

d) Como puede verse, las técnicas anteriormente descritas son combinables ya que el estudio de un conjunto de datos cualitativos puede comenzar con un «análisis de correspondencias» para obtener coordenadas (cuantitativas) en los nuevos ejes factoriales que, tras una reducción de dimensiones mediante un «análisis de compontes principales» pueden ser, en un tercer paso, agrupadas en clases a través de un «análisis de conglomerados» y así sucesivamente. Por este motivo, resulta interesante analizar

[11] Debe definirse como va a medirse esta «proximidad», lo cual da origen a diferentes variantes de la técnica.

la forma de concatenar las diferentes operaciones de forma que se formen procedimientos de análisis más complejos y, en especial, resulta de interés recurrir a sistemas informáticos que permitan diseñarlos utilizando una interfaz gráfica.

4.- Técnicas de clasificación

El presente apartado consta de dos partes. En la primera se describen las fichas que se han realizado de cada técnica con el objeto de comprender su fundamente teórico y ver su aplicación práctica al tratamiento de datos, particularizado a las bases de datos que se están utilizando (medidas de dientes de dinosaurios). En segundo lugar, se presentan una serie de reflexiones sobre el empleo de estas técnicas que, posteriormente, servirán para la confección de la propuesta de procesamiento de los datos.

4.1.- Catálogo de técnicas analizadas

Como se ha indicado en el apartado anterior, se han seleccionado cuatro técnicas para su análisis, teniendo en cuenta su uso actual en las tareas de clasificación, así como su pertinencia para abordar diferentes características específicas del problema que se está estudiando (como el hecho de tratar atributos «cualitativos»).

El proceso para cada técnica parte de la realización de un pequeño documento inicial en el cual se explica brevemente el fundamento de la técnica y se presenta la formulación práctica de manera apropiada para su implementación (sin entrar en el detalle de las pruebas ni los desarrollos que justifican las expresiones correspondientes). Los documentos también incluyen un ejemplo desarrollado con los datos paleontológicos de prueba suministrados que permiten ilustrar la aplicación en el contexto del proyecto que se está realizando. La realización de estos ejemplos se efectúa con diversas herramientas (hoja de cálculo, software específico[12] y/o mediante el desarrollo de aplicaciones informáticas[13]).

Además de la secuencia de cálculos, se ha prestado especial atención a las características que deben cumplir los datos para asegurar que los resultados que se obtengan sean consistentes con la teoría que los sustenta. Aspectos como si los valores de las diferentes variables deben expresarse o no en las mismas unidades, requieren o no ser normalizados, etc. Así mismo, se ha procurado comprender el significado y representatividad de los diferentes tipos de resultados.

El listado concreto de técnicas analizadas ha incluido las siguientes:

1) T01.- Análisis de componentes principales.
2) T02.- Análisis de correspondencias.
3) T03.- Análisis factorial discriminante.
4) T04.- Análisis de conglomerados (clúster).

Cada uno de estos documentos ha partido de una redacción inicial que posteriormente ha sido revisada y completada por el resto de participantes en esta línea del proyecto de investigación para generar la versión definitiva. Dado el marco temporal del presente proyecto, estas

[12] En concreto se ha utilizado *Wolfram Mathematica*® para el cálculo de autovalores y autovectores.

[13] Por ejemplo, en el caso del caso del «análisis clúster» ya que una gran parte del algoritmo se basa en el manejo recursivo de una lista de elementos cuya longitud va reduciéndose sucesivamente según avanzan las iteraciones. Para realizar las pruebas, se confeccionó un programa específico en lenguaje *Tcl*.

segundas versiones son las últimas que se han generado en cada caso, si bien se debe considerar que se trata de documentos abiertos que serían susceptibles de ser aumentados en el futuro, con exposiciones más detalladas y que abarcasen aspectos que no se han podido tratar en estas primeras aproximaciones a las técnicas, otros tipos de resultados obtenibles y extensiones.

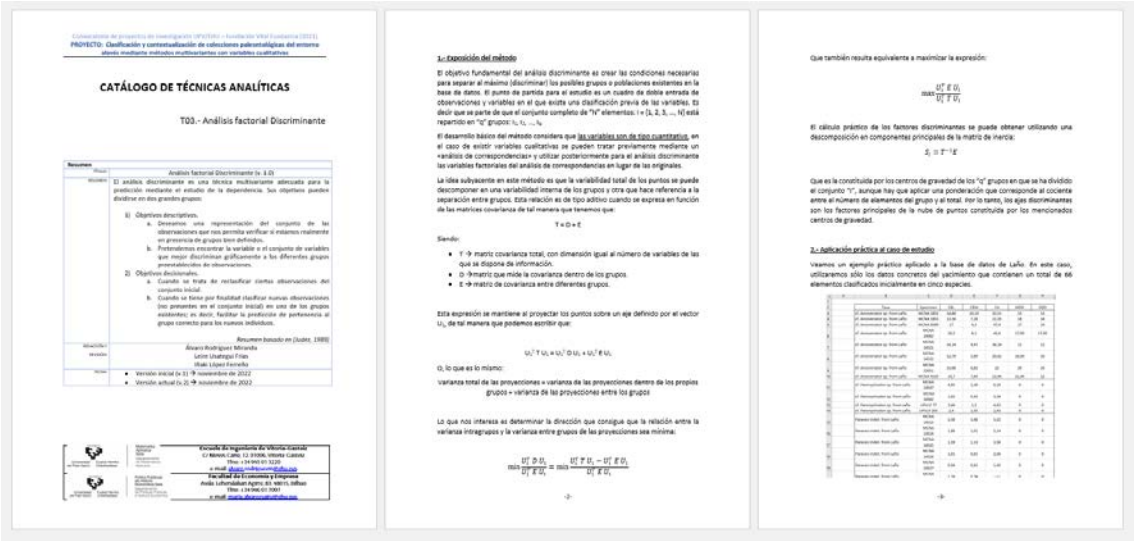


Fig. 8.- Primeras hojas del documento relativo a la técnica del «análisis discriminante».

Por otro lado, la lista de técnicas podría ampliarse en el futuro con nuevas herramientas como la regresión, los árboles de decisión, etc. que permitirían ir enriqueciendo las posibilidades de análisis[14]. Como se verá más adelante, el entorno gráfico que se ha empleado para crear de manera visual los organigramas de procesado de la información (Orange) cuenta con numerosos controles que permiten aplicar muchas de estas técnicas por lo que son un buen punto de partida para ir seleccionándolas y estudiándolas, de manera que una vez que se comprenda el alance y uso de cada una se pueda integrar directamente en los flujos de trabajo.

Estos documentos se consideran material interno, es decir, que no están preparados para su difusión (la forma de redactar los textos, referencias bibliográficas, compleción de las fórmulas, etc., no son adecuadas para ello). Como se ha indicado, son documentos de trabajo cuyo interés es la formación a través de la práctica y la compartición de conocimiento entre los miembros del proyecto que tienen que trabajar con estas técnicas. Por este motivo, se presentan como anexos en la versión que se entrega como resultados del proyecto, pero no se incluirán en la

[14] Respecto a las "nuevas" piezas que se pueden proponer, es interesante mencionar el «análisis log-ratio» que es especialmente adecuado para casos en que todas las variables se miden en una misma unidad y los valores son estrictamente positivos. En este caso, la matriz sobre la que se obtiene la reducción de dimensiones está formada por logaritmos, pero existe un doble centrado (tanto por filas como por columnas) de tal manera que las proporciones que se miden a través de los logaritmos se pueden interpretar tanto dentro de una misma variable como entre los valores obtenidos en diferentes variables. Precisamente, los análisis log-ratio son empleados en casos de morfometría muy similares al aquí analizado de los dientes de dinosaurio. Como referencia se remite al trabajo siguiente:

- Greenacre, M. (2010) *Biplots in practice*. Fundación BBVA.

versión que estará disponible para su difusión pública a través del repositorio de la universidad[15].

4.2.- Consideraciones sobre la aplicación práctica de las técnicas

Existe una característica común a las diferentes herramientas de clasificación que es la «normalización de los datos». En efecto, es una situación frecuente que se cuente con información multivariante que se haya podido medir en diferentes unidades y/o que corresponda a variables que aun midiéndose en una unidad común tengan rangos de variación muy diferentes. Las distancias euclídeas —que se utilizan para determinar cómo de similares o separados se encuentran dos elementos concretos— en las cuales todas las variables intervienen de la misma manera pueden dar resultados inapropiados (e incluso absurdos) si las variables son heterogéneas y entran tal cual fueron medidas originalmente en el cálculo. Por ello, un paso previo a la aplicación de los algoritmos de clasificación suele ser la de hacer todas las variables comparables, una de las formas más habituales consiste en recurrir a variables «estandarizadas», es decir, aquellas en las que a los valores originales (x_i) se les resta sus respectivas medias (x_m) y se dividen por la desviación típica (s_x)[16].

$$z_i = \frac{x_i - x_m}{s_x}$$

De esta manera, la variable resultante (z_i) pasa a ser adimensional y los valores se miden en número de desviaciones típicas respecto a la correspondiente media. Por ejemplo, las siguientes tablas muestran este proceso realizado sobre los datos de la base de datos del yacimiento de prueba de Lo Hueco, en el cual se dispone de cinco valores numéricos (variables: x_1, x_2, x_3, x_4 y x_5) de un conjunto de especímenes de dientes. En primer lugar, se calculan las medias y desviaciones típicas de cada columna (variable).

[15] Disponible en: http://hdl.handle.net/10810/58547

[16] En todo caso, este paso no debe aplicarse de forma sistemática sin una consideración previa ya que —dentro de un determinado contexto— diferentes variables pueden ser comparables, aunque sus respectivas medias y desviaciones típicas varíen. Por ejemplo, si disponemos de una tabla en la cual para un conjunto de países (observaciones) tenemos los salarios de dos tipos de profesiones (variables) por ejemplo (ingenieros y médicos), si estandarizamos las variables estaremos analizando variaciones con respecto a las situaciones promedio de cada profesión; mientras que si utilizamos los datos brutos se estarán haciendo comparaciones directas sobre los sueldos (por ejemplo, en euros), al tratarse en ambos casos de salarios, este segundo tipo de análisis también resulta posible.

Clasificacion	Sigla nueva	x1 altura total	x2 FABL	x3 anchura	x4 dd/mm	x5 dm/mm
Dromi	LOH-ter- 1	8	7	3,33	5	6
Dromi	LOH-ter- 2	6	3,87	2,33	3	5,5
Velocir	LOH-ter- 3	6,07	4,2	2	3	4
Velocir	LOH-ter- 4	11,33	6,87	3,67	3	3,5
Velocir	LOH-ter- 5	14,95	6	3	3	4,5
Velocir	LOH-ter- 6	9,53	4,47	3	2,5	0
veloc	LOH-ter-141	1,6	0,72	0,45	3	3,5
veloc	LOH-ter-142	1,2	0,5	0,3	3,5	0
veloc	LOH-ter-143	1,61	0,84	0,43	3	3
cf. Pyrop	LOH-ter-144	1,04	0,49	0,25	6	8
dromi	LOH-ter-145	1,1	0,54	0,32	3	0
	media (xm):	8,56	4,34	2,37	3,25	1,54
	desv. Típica (sx):	5,11	2,29	1,34	0,85	2,27
Número de elementos (N):		128				

Fig. 9.- Valores de las medias y desviaciones típicas de cada una de las variables cuantitativas.

Para, seguidamente, proceder a estandarizar todas estas variables, obteniendo las correspondientes variables (z), que son las que se utilizarán en los análisis posteriores.

		x1	x2	x3	x4	x5	Variables centradas y reducidas				
Clasificacion	Sigla nueva	altura total	FABL	anchura	dd/mm	dm/mm	z1	z2	z3	z4	z5
Dromi	LOH-ter- 1	8	7	3,33	5	6	-0,11	1,16	0,71	2,05	1,97
Dromi	LOH-ter- 2	6	3,87	2,33	3	5,5	-0,50	-0,20	-0,03	-0,30	1,75
Velocir	LOH-ter- 3	6,07	4,2	2	3	4	-0,49	-0,06	-0,28	-0,30	1,08
Velocir	LOH-ter- 4	11,33	6,87	3,67	3	3,5	0,54	1,11	0,97	-0,30	0,86
Velocir	LOH-ter- 5	14,95	6	3	3	4,5	1,25	0,73	0,47	-0,30	1,31
Velocir	LOH-ter- 6	9,53	4,47	3	2,5	0	0,19	0,06	0,47	-0,89	-0,68
Velocir	LOH-ter- 7	8,4	4,67	2,67	2,5	0	-0,03	0,15	0,22	-0,89	-0,68
Velocir	LOH-ter- 8	13	6	3	3	3	0,87	0,73	0,47	-0,30	0,64

Fig. 10.- Valores estandarizados de las variables cuantitativas.

En todo caso, la «estandarización» de las variables no es la única alternativa de hacerlas comparables, de hecho, cuando todas las variables consisten en diferentes medidas realizadas en una misma unidad, una opción habitual es tomar logaritmos (Greenacre, 2010). De esta manera, si inicialmente teníamos dos variables ("a" y "b", imaginemos que se relacionan con el "alto" y el "ancho" de un objeto) que, para dos elementos (i_1 e i_2) tienen sus correspondientes valores: a_{i1}, b_{i1}, a_{i2} y b_{i2}. La diferencia directa entre ellos: ($a_{i1} - b_{i1}$) frente a ($a_{i2} - b_{i2}$), está bien definida en cuanto a unidades, pero puede carecer de significado práctico (por ejemplo, en el caso de que se consideren objetos de tamaños muy diferentes). Sin embargo, la diferencia entre los logaritmos cumple con la siguiente propiedad:

$$\text{Ln}\,(a_{i1}) - \text{Ln}\,(b_{i1}) = \text{Ln}\,(a_{i1} / b_{i1})$$

$$\text{Ln}\,(a_{i2}) - \text{Ln}\,(b_{i2}) = \text{Ln}\,(a_{i2} / b_{i2})$$

Lo que implica que lo que realmente se están analizando son las proporciones ("alto" dividido entre el "ancho") y éstas sí que pueden ser comparables incluso entre especímenes con tamaños muy diferentes.

Ésta es, precisamente, la situación que se presenta en el problema de la morfología de los dientes. De hecho, la «normalización» que se realiza en el trabajo sobre los dientes de Laño utiliza el logaritmo de las variables originales con una consideración adicional que pasamos a comentar.

De todas las medidas que se han medido para los dientes de Laño (hasta 27 en algunos casos), la base de datos de referencia (dientes de otros yacimientos ya clasificados) sólo cuenta con cinco de estas medidas, en concreto: CBL, CBW, CH, MDD, DDD, por lo que los análisis posteriores se limitan sólo a estas cinco. Las tres primeras corresponden con las dimensiones principales (longitud, anchura y altura) medidas en milímetros, mientras que las dos últimas son densidades de dentículos en los filos anterior y posterior de los dientes (en concreto, el número de dentículos en un tramo de 5 mm). Como se ha indicado, la normalización de estas variables se realiza a través de sus respectivos logaritmos; pero dado que las variables relativas a las densidades de dentículos pueden ser iguales a cero (en el caso de que los dientes no presenten dentículos) y que el logaritmo de cero no está definido, esta operación no es aplicable. Como solución, se recurre a sumar una unidad a la variable antes de obtener su logaritmo.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1				Raw data					Normalized data				
2	Taxa	Reference	Specimen	CBL	CBW	CH	MDD	DDD	CBL	CBW	CH	MDD	DDD
3	*Pyroraptor olympius*	RonanAllain pers. comm.	?	6,1	3,1	16	42,5	30	0,851258349	0,612783857	1,230448921	1,638489257	1,491361694
4	cf. *Arcovenator* Armuña	Pérez-García et al., 2016	UPUAM 14044	21,3	11,4	45,5	15	15	1,348304863	1,093421685	1,667452953	1,204119983	1,204119983
5	cf. *Arcovenator* Armuña	Pérez-García et al., 2016	UPUAM 14047	21	9,2	41	15	15	1,342422681	1,008600172	1,62324929	1,204119983	1,204119983
6	cf. *Arcovenator* Armuña	Pérez-García et al., 2016	UPUAM 14048	14	6,4	19,1	15	15	1,176091259	0,86923172	1,303196057	1,204119983	1,204119983
7	*Arcovenator escotae*	Tortosa et al., 2014	MHNAIX-PV-2011-12-15	15	8	52,65	17,5	20	1,204119983	0,954242509	1,729569726	1,267171728	1,322219295
8	*Arcovenator escotae*	Tortosa et al., 2014	MHNAIX-PV-2011-12-187	19	10	51	17,5	17,5	1,301029996	1,041392685	1,716003344	1,267171728	1,267171728
9	*Majungasaurus*	Smith et al., 2005	FMNHPR2008	12,99	9,46	30,11	11	11,5	1,145817714	1,019531685	1,492900011	1,079181246	1,096910013
10	*Majungasaurus*	Smith et al., 2005	UA 8716	12,41	9,26	27,05	10	10	1,127428778	1,011147361	1,447932866	1,041392685	1,041392685
11	*Majungasaurus*	Smith et al., 2005	UA 8716	12,51	8,3	27,69	10	11	1,130655349	0,968482949	1,457730548	1,041392685	1,079181246
12	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,3	8,62	36,9	8,5	9	1,285557309	0,983175072	1,57863921	0,977723605	1
13	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,37	9,21	38,08	9,5	9	1,287129621	1,009025742	1,591954555	1,021189299	1
14	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,93	8,86	35,54	10	10	1,299507299	0,993876915	1,562768543	1,041392685	1,041392685
15	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	18,17	9,1	38,68	9,7	9	1,282622113	1,004321374	1,598571663	1,029383778	1
16	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	7,88	3,47	12,45	12	14	0,948412966	0,650307523	1,128722284	1,113943352	1,176091259
17	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	8,81	7,2	19,88	9	9	0,991669007	0,913813852	1,319730494	1	1
18	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	13,3	8,56	25,37	11	9,5	1,155336037	0,980457892	1,42111013	1,079181246	1,021189299
19	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	14,24	7,72	25,13	11	11	1,182984967	0,940516485	1,41713941	1,079181246	1,079181246
20	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,66	7,12	19,93	12	10	1,135450699	0,909556029	1,320769228	1,113943352	1,041392685
21	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,53	6,69	19,21	11,7	10,5	1,131297797	0,88592634	1,305566314	1,103803721	1,06069784
22	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	12,28	6,7	17,87	12	11,3	1,123198075	0,886490725	1,2757719	1,113943352	1,089905111
23	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	11,72	6,29	16,19	12,8	10,5	1,104487111	0,862727528	1,235275877	1,139879086	1,06069784
24	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	9,33	5,36	14,48	15	11	1,014100322	0,803457116	1,189770956	1,204119983	1,079181246
25	*Majungasaurus*	Smith et al., 2005	FMNH PR2100	10,91	8,48	22,88	8	8,5	1,075911761	0,976808337	1,378034322	0,954242509	0,977723605

Fig. 11.- Base de datos de referencia, valores originales de las variables seleccionadas (x) y valores normalizados: z = ln (x + 1) (Isasmendi *et al.*, 2022).

Con el fin de poder seguir con el hilo argumental, se ha rehecho el cálculo de los componentes principales tal como se indica en el artículo que se ha realizado hasta la representación gráfica bidimensional de las coordenadas correspondientes a los especímenes de los dientes en los dos ejes factoriales principales.

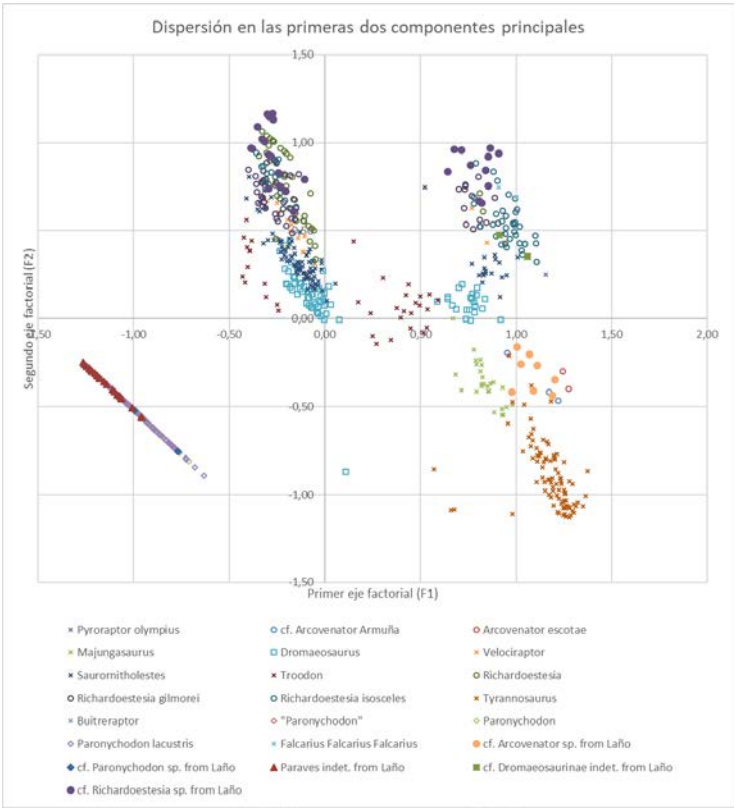


Fig. 12.- Representación de los elementos según las coordenadas en los dos primeros ejes obtenidos en el análisis de componentes principales —según los datos y el procedimiento indicado en (Isasmendi *et al.*, 2022)—; los símbolos sólidos corresponden a los ejemplares del yacimiento de Laño, los símbolos similares vacíos corresponden a especímenes de la misma especie de la base de referencia y los que aparecen señalados con una equis son otros tipos de dinosaurios (no entramos, por el momento, a discutir sobre la similitud de éstos con los presentes en Laño).

A grandes rasgos, la interpretación general de este gráfico parece estar marcada por una dirección relativa al tamaño de los dientes y otra que indica el número de dentículos. A modo ilustrativo, sobre el gráfico anterior se han seleccionado algunos grupos significativos sobre los que se indican, aproximadamente, el rango de variación de las variables CH (altura)[17] y DDD (número de dentículos en un tramo de 5 mm en el filo trasero del diente):

[17] En el artículo (Isasmendi *et al.*, 2022) se indica que la explicación del primer eje principal es debida, en mayor medida, al tamaño (CH, CBL y CBW) y la segunda a la densidad de los dentículos (en especial, DDD). Asimismo, se indica que las dos primeras componentes explican respectivamente el 64% y el 22,6% de la inercia total (es decir, que entre ambas representan el 86,6% de la varianza). En el cálculo repetido por nosotros, los valores de inercia explicada que obtuvimos fueron ligeramente diferentes (52% y 30% para un total del 82%).

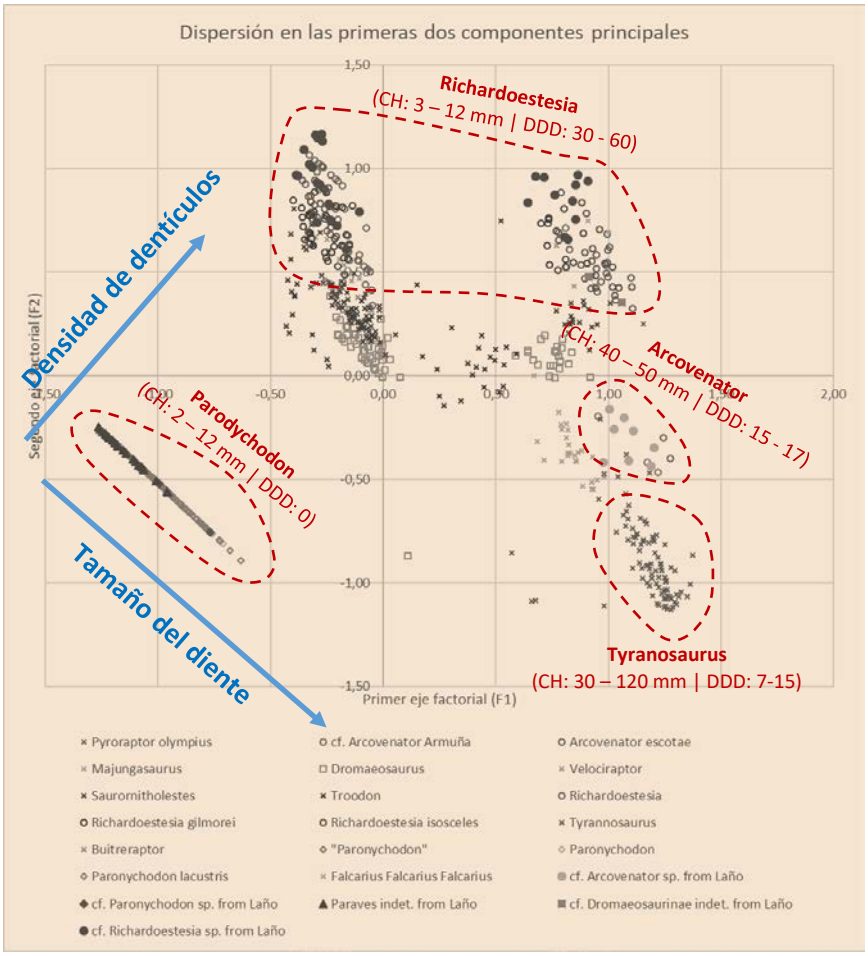


Fig. 13.- Interpretación de las variables explicativas de la representación de los especímenes sobre el plano formado por los dos primeros ejes factoriales del «análisis de componentes principales».

El siguiente paso, consiste en la interpretación paleontológica del gráfico sobre la que, por el momento, no vamos a profundizar ya que nos interesa más continuar con la forma en que se han presentado los valores numéricos para proceder al cálculo[18].

En efecto, dado que ésta es una forma habitual de «normalizar» los datos en el contexto del presente problema, hemos de suponer que los resultados que ofrece son (al menos aproximadamente) aceptables. No obstante, si recordamos el fundamento matemático de

[18] Debe tenerse en cuenta que los especímenes ya cuentan con una clasificación (de tipo «experto») que indican a qué especie pertenecen, la representación gráfica permite validar hipótesis o plantear posibles respuestas a las aparentes incongruencias. Es decir, que en un mismo gráfico como el que se ha presentado se pueden encontrar situaciones como la del *Arcovenator* en la que la nube de puntos de los especímenes del yacimiento de Laño se superpone bien con la de los especímenes de referencia (lo cual parece sugerir que la identificación es correcta), la del *Dromaeosaurus* en que las nubes de puntos de los especímenes de Laño y de los de referencia son ambas densas pero están separadas (lo cual podría justificarse por el hecho de que se trate de dos subespecies diferentes —los de referencia son especímenes americanos y los de Laño son europeos—) o los *Richardoestesia* que, bajo la misma denominación parecen formar dos conjuntos separados sobre el gráfico (características que ha llevado a suponer la existencia de variabilidad dentro de la dentición de este animal). En definitiva, que no se debe esperar que el gráfico identifique de manera clara todas las especies, al contrario, el gráfico en sí muestra una situación que en algunos casos se corresponderá y en otros no con la clasificación previa, tanto las coincidencias como las divergencias son relevantes y permiten fundamentar la interpretación que posteriormente se haga.

haber utilizado logaritmos (es decir, poder comparar las proporciones) nos encontramos con que el hecho de sumar una unidad lo distorsiona; además, las tres primeras variables (CBL, CBW y CH) no lo necesitan.

Como alternativa, se propone, mantener las variables originales (de tipo "x") correspondientes a las distancias medidas en milímetros (CBL, CBW, CH) y reemplazar las densidades (MDD, DDD) por los tamaños de los dentículos (nuevamente en milímetros), es decir, utilizar unas nuevas variables originales que sean:

$$x_4 = 5 / MDD \quad y \quad x_5 = 5 / DDD$$

En el caso de que los denominadores sean iguales a cero (ausencia de dentículos) se recurre a hacer que las variables anteriores (x_4 y x_5) tomen el valor de la longitud total del diente[19] (CH).

A continuación, se consideran las variables «normalizadas» (z_i) como los logaritmos de las correspondientes variables originales (x_i).

	Normalización alternativa					Variables normalizadas: z = ln (x)				
Taxa	x1 = CBL	x2 = CBW	x3 = CH	x4 = 5/ MDD	x5 = 5 / DDD	z1	z2	z3	z4	z5
Pyroraptor olympius	6,1	3,1	16	0,118	0,167	0,785	0,491	1,204	-0,929	-0,778
cf. Arcovenator Armuña	21,3	11,4	45,5	0,333	0,333	1,328	1,057	1,658	-0,477	-0,477
cf. Arcovenator Armuña	21	9,2	41	0,333	0,333	1,322	0,964	1,613	-0,477	-0,477
cf. Arcovenator Armuña	14	6,4	19,1	0,333	0,333	1,146	0,806	1,281	-0,477	-0,477
Arcovenator escotae	15	8	52,65	0,286	0,250	1,176	0,903	1,721	-0,544	-0,602
Arcovenator escotae	19	10	51	0,286	0,286	1,279	1,000	1,708	-0,544	-0,544
Majungasaurus	12,99	9,46	30,11	0,455	0,435	1,114	0,976	1,479	-0,342	-0,362
Majungasaurus	12,41	9,26	27,05	0,500	0,500	1,094	0,967	1,432	-0,301	-0,301
Majungasaurus	12,51	8,3	27,69	0,500	0,455	1,097	0,919	1,442	-0,301	-0,342
Majungasaurus	18,3	8,62	36,9	0,588	0,556	1,262	0,936	1,567	-0,230	-0,255
Majungasaurus	18,37	9,21	38,08	0,526	0,556	1,264	0,964	1,581	-0,279	-0,255
Majungasaurus	18,93	8,86	35,54	0,500	0,500	1,277	0,947	1,551	-0,301	-0,301
Majungasaurus	18,17	9,1	38,68	0,515	0,556	1,259	0,959	1,587	-0,288	-0,255

Fig. 14.- Tabla de datos con las variables (X) expresadas todas en milímetros y (Z) con sus respectivos logaritmos.

Antes de calcular las componentes principales, se restan las medias de cada columna de tal manera que las variables queden centradas.

[19] Es decir, en el caso de que la morfología no presente dentículos se supone que todo el filo corresponde a un único dentículo. Esta forma de considerar la variable resuelve el problema del valor cero en la cuenta directa de dentículos (que ofrece problemas posteriores si se quieren tomar logaritmos) a la vez que presenta un valor para la situación sin dentículos muy diferente a las situaciones que sí que muestran dentículos… no obstante, la interpretación de este hecho desde el punto de vista morfológico-evolutivo no resulta satisfactoria ya que los dientes sin dentículos no corresponden a un caso límite de extensión del tamaño hasta llegar a alcanzar toda la longitud del diente, sino que es el resultado de un proceso evolutivo que empieza con un conjunto regular de dentículos (con independencia de su densidad) —que es la situación de los terópodos— que pasa a ser una disposición irregular —en el caso de los cocodrilos— a una siguiente etapa en que ya no existen dentículos como tales sino irregularidades del filo pasando, finalmente, a una situación en la que los dientes son lisos.

							Restamos los promedios de cada columna				
Taxa	z1	z2	z3	z4	z5		z1'	z2'	z3'	z4'	z5'
Pyroraptor olympius	0,785	0,491	1,204	-0,929	-0,778		0,118	0,098	0,299	-1,224	-0,479
cf. Arcovenator Armuña	1,328	1,057	1,658	-0,477	-0,477		0,661	0,664	0,753	-0,772	-0,178
cf. Arcovenator Armuña	1,322	0,964	1,613	-0,477	-0,477		0,655	0,571	0,708	-0,772	-0,178
cf. Arcovenator Armuña	1,146	0,806	1,281	-0,477	-0,477		0,479	0,413	0,376	-0,772	-0,178
Arcovenator escotae	1,176	0,903	1,721	-0,544	-0,602		0,509	0,510	0,817	-0,838	-0,303
Arcovenator escotae	1,279	1,000	1,708	-0,544	-0,544		0,612	0,607	0,803	-0,838	-0,245
Majungasaurus	1,114	0,976	1,479	-0,342	-0,362		0,447	0,583	0,574	-0,637	-0,062
Majungasaurus	1,094	0,967	1,432	-0,301	-0,301		0,427	0,573	0,527	-0,595	-0,002
Majungasaurus	1,097	0,919	1,442	-0,301	-0,342		0,430	0,526	0,538	-0,595	-0,043
Majungasaurus	1,262	0,936	1,567	-0,230	-0,255		0,595	0,542	0,662	-0,525	0,044
Majungasaurus	1,264	0,964	1,581	-0,279	-0,255		0,597	0,571	0,676	-0,573	0,044
Majungasaurus	1,277	0,947	1,551	-0,301	-0,301		0,610	0,554	0,646	-0,595	-0,002
Majungasaurus	1,259	0,959	1,587	-0,288	-0,255		0,592	0,566	0,683	-0,582	0,044
cf. Richardoestesia sp. from Laño	0,486	0,130	0,583	-0,903	-0,845		-0,181	-0,263	-0,322	-1,198	-0,546
cf. Richardoestesia sp. from Laño	0,083	-0,260	0,196	-1,041	-0,903		-0,584	-0,653	-0,709	-1,336	-0,604
cf. Richardoestesia sp. from Laño	0,201	-0,187	0,444	0,444	-1,041		-0,466	-0,580	-0,461	0,150	-0,742
cf. Richardoestesia sp. from Laño	0,155	-0,237	0,524	-1,079	-1,041		-0,512	-0,630	-0,381	-1,374	-0,742
cf. Richardoestesia sp. from Laño	0,161	-0,125	0,509	-1,079	-1,000		-0,506	-0,518	-0,396	-1,374	-0,701
cf. Richardoestesia sp. from Laño	0,076	-0,268	0,365	0,365	-1,079		-0,591	-0,661	-0,539	0,071	-0,780
cf. Richardoestesia sp. from Laño	0,170	-0,086	0,356	0,356	-1,041		-0,497	-0,479	-0,549	0,062	-0,742
medias:	0,667	0,393	0,905	0,294	-0,299		0,000	0,000	0,000	0,000	0,000
Número de elementos (N):	759										

Fig. 15.- Variables (Z') que corresponden a los logaritmos centrados, de forma que las medias respectivas de cada columna sean iguales a cero.

El cálculo de los componentes principales sobre este conjunto de datos, ofrece el siguiente resultado.

	Valor prop.					
% explicado	λ	Vector propio				
0,47	0,632743	-0,4500229	-0,4990378	-0,4680765	0,5367535	0,203078
0,31	0,420071	0,2705076	0,3288677	0,3433778	0,6186189	0,5639805
0,21	0,280454	-0,095156	-0,0699854	-0,1421147	-0,5713225	0,7996508
0,01	0,0090241	-0,1400024	-0,6308757	0,760982	-0,0506057	0,0272126
0,00	0,0034389	0,8340565	-0,4898023	-0,2523984	0,0152988	0,022457
suma:	1,35					
De los que seleccionamos los dos primeros para la representación gráfica.						
El porcentaje de inercia explicado por los dos primeros valores propios es de:						
	PHE2 =	0,78				

Fig. 16.- Valores y vectores propios obtenidos a partir de los datos (variables Z'). En la parte inferior, se muestra que el porcentaje de la inercia explicado por las dos primeras componentes es del 78%.

La representación gráfica de las coordenadas proyectadas en los dos ejes principales indicados es la que se muestra a continuación:

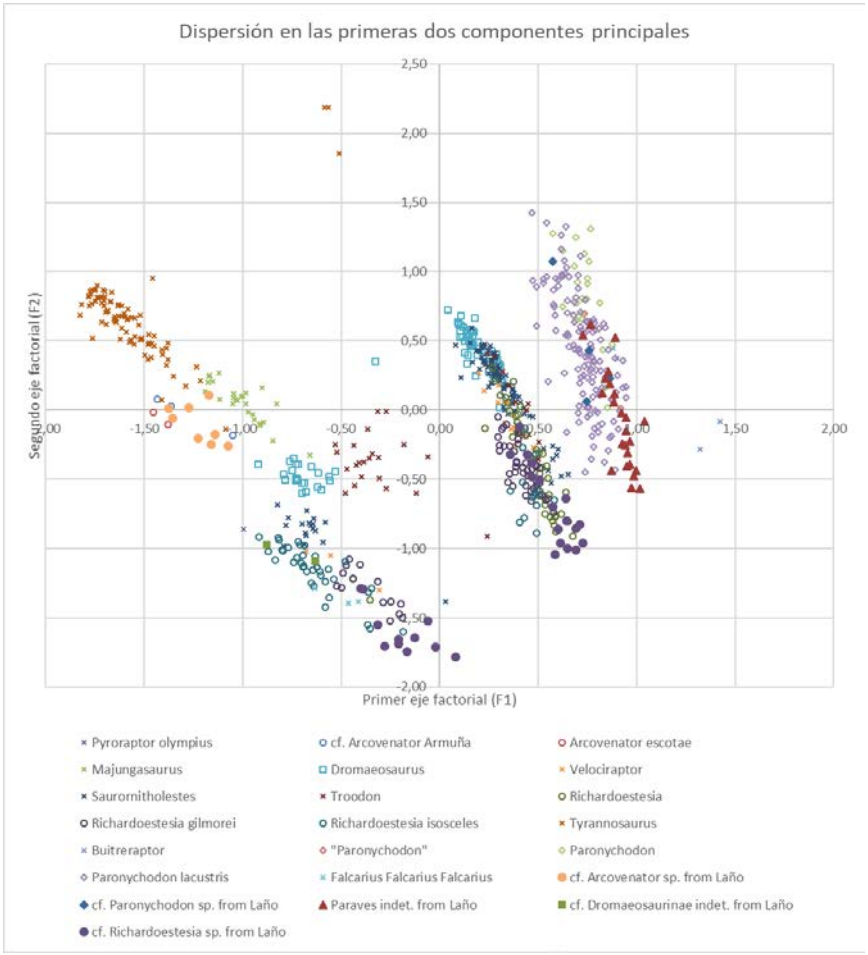


Fig. 17.- Representación de los elementos según las coordenadas en los dos primeros ejes obtenidos en el análisis de componentes principales sobre las variables normalizadas según la propuesta indicada. Los símbolos sólidos corresponden a los ejemplares del yacimiento de Laño, los símbolos similares vacíos corresponden a especímenes de la misma especie de la base de referencia y los que aparecen señalados con una equis son otros tipos de dinosaurios.

El gráfico es similar al obtenido inicialmente en el artículo de Isasmendi *et al*. (2022) que se ha presentado anteriormente (aunque con un giro de 180º con respecto al centro que no afecta al objeto del análisis que es mostrar la máxima dispersión posible de los elementos).

Al igual que se ha hecho antes, se muestran sobre el gráfico algunas de las familias de especímenes que permiten ver la capacidad de separar entre los diferentes tipos de dientes.

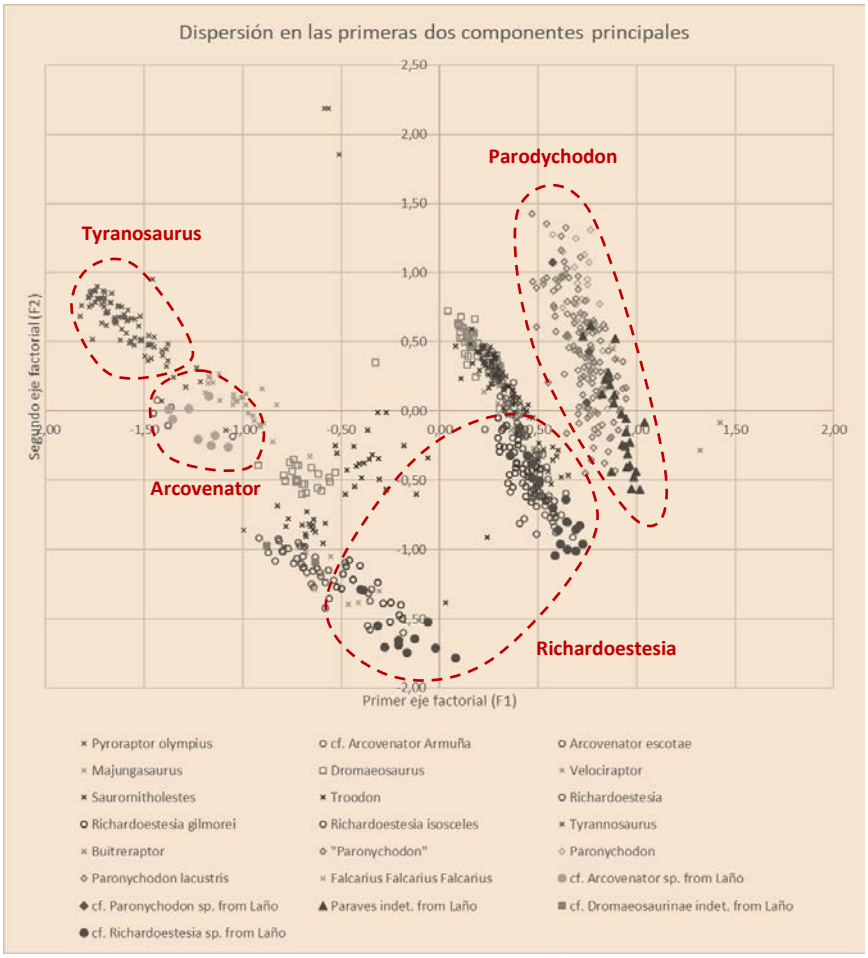


Fig. 18.- Agrupación de los especímenes de las cuatro especies indicadas en el análisis original por componentes principales. A diferencia del gráfico anterior, se aprecia que los especímenes de *Parodychodon* aparecen distribuidos bidimensionalmente (ganando así una dimensión respecto al análisis del artículo original).

Se puede continuar el cálculo, determinando las correlaciones entre las variables (Z') y las transformadas (f_1 y f_2, que son las direcciones principales).

Correlaciones con las variables en los ejes factoriales					
	z1	z2	z3	z4	z5
f1	-0,88	-0,87	-0,83	0,65	0,28
f2	0,43	0,47	0,50	0,61	0,63

Fig. 19.- Correlaciones entre las variables y los dos primeros ejes factoriales.

Si volvemos la atención a los vectores propios que se utilizan para obtener las coordenadas en los ejes factoriales, tenemos que las coordenadas en los nuevos ejes son:

$$f_1 = -0{,}45\ z_1' - 0{,}50\ z_2' - 0{,}47\ z_3' + 0{,}54\ z_4' + 0'2\ z_5' \approx -0{,}5 \ln(x_1 \cdot x_2 \cdot x_3) + 0{,}5 \ln(x_4) + 0{,}2 \ln(x5) + C_1$$

$$f_2 = 0{,}27\ z_1' + 0{,}32\ z_2' + 0{,}34\ z_3' + 0{,}61\ z_4' + 0'56\ z_5' \approx 0{,}3 \ln(x_1 \cdot x_2 \cdot x_3) + 0{,}6 \ln(x_4 \cdot x_5) + C_2$$

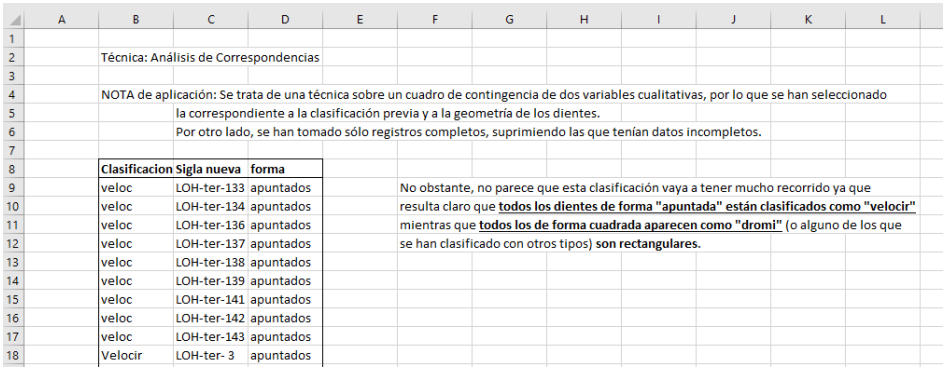
En ambos casos se constata que aparece el término ($x_1 \cdot x_2 \cdot x_3$) es decir (CBL · CBW · CH) o, lo que es lo mismo, la referencia al volumen del diente. Asimismo, también se presenta la influencia de las medidas correspondientes al tamaño de los dentículos.

El tercer factor principal (no representado en el gráfico 2D) parece estar relacionada con el cociente: x_5/x_4 (relación entre los tamaños de los dentículos anteriores y posteriores).

Por supuesto, estas consideraciones sobre la interpretación de los ejes resultantes de la transformación se realizan partiendo sólo del ejemplo que se ha trabajado aquí y sería necesario comprobar si se vuelven a presentar en otras bases de datos y si tienen sentido en el ámbito de la morfología de los dientes.

En el tema de los dentículos existen dos opciones, la primera es considerarlos como variables de naturaleza cuantitativa, en los cuales el hecho de que no existan dentículos es una situación más en un continuo de posibilidades que incluye las opciones de: 0 dentículos, 1 dentículo, 2 dentículos, etc. Frente a esta opción, también se puede considerar que se trata de una variable categórica y que, por lo tanto, el hecho de que no existan dentículos debe tratarse como una situación completamente diferente al hecho de que sí que los haya (con independencia de que sean muchos o pocos). El enfoque utilizado en el artículo original es del primer tipo (considerar a los dentículos como una variable «cuantitativa») y también en esta línea va la propuesta alternativa que se ha presentado (en la que, en vez del número de dentículos, lo que se analiza es el tamaño de los mismos). Igualmente en esta línea podrían situarse los modelos lineales generalizados (Greenacre, 2010, capítulo 3) para el caso de la regresión de Poisson y logístico, ambos tienen interés ya que permiten considerar el cero como valor válido, sin embargo, quizás no sean del todo adecuados ya que el primero —Poisson— considera el cero como uno de los resultados posibles en una cuenta (0, 1, 2, 3… dentículos) lo cual considera todos estos casos posibles dentro del mismo elemento (y no el hecho de que exista una clara diferencia entre los dientes con y sin dentículos), además, sólo considera valores enteros. El segundo caso —logística— sí que refleja bien el hecho de que existan o no dentículos (ya que los posibles valores son uno o cero) pero no permite diferenciar en función de la densidad concreta de dentículos dentro de los casos en que éstos existan.

Por el contrario, el enfoque de considerar los dentículos como variable categórica se puede abordar a través del «análisis de correspondencias». No obstante, las bases de datos disponibles no han permitido analizar este punto con detalle, por ejemplo, en la suministrada sobre el yacimiento de Lo Hueco resultaba que todos los dientes preclasificados como «Velociraptor» tenían los dentículos con forma apuntada mientras que los identificados como «Dromeosaurio» tenían todos los dentículos rectangulares, de esta manera, la tabla de correspondencias resultaba trivial.



Técnica: Análisis de Correspondencias

NOTA de aplicación: Se trata de una técnica sobre un cuadro de contingencia de dos variables cualitativas, por lo que se han seleccionado la correspondiente a la clasificación previa y a la geometría de los dientes.
Por otro lado, se han tomado sólo registros completos, suprimiendo las que tenían datos incompletos.

Clasificacion	Sigla nueva	forma
veloc	LOH-ter-133	apuntados
veloc	LOH-ter-134	apuntados
veloc	LOH-ter-136	apuntados
veloc	LOH-ter-137	apuntados
veloc	LOH-ter-138	apuntados
veloc	LOH-ter-139	apuntados
veloc	LOH-ter-141	apuntados
veloc	LOH-ter-142	apuntados
veloc	LOH-ter-143	apuntados
Velocir	LOH-ter- 3	apuntados

No obstante, no parece que esta clasificación vaya a tener mucho recorrido ya que resulta claro que **todos los dientes de forma "apuntada" están clasificados como "velocir"** mientras que **todos los de forma cuadrada aparecen como "dromi"** (o alguno de los que se han clasificado con otros tipos) **son rectangulares.**

Fig. 20.- Intento de análisis de los datos del yacimiento de Lo Hueco a través del «análisis de correspondencias».

En otro orden de ideas, resulta interesante considerar la opción de que quizás no todos los dientes deben analizarse con el mismo conjunto de atributos. Esto resulta especialmente patente viendo que las medidas disponibles en los diferentes especímenes son bastante diversas y que, al tener que recurrir sólo a las medidas comunes para el análisis conjunto, la riqueza de la caracterización disponible se desaprovecha en gran medida.

Por otro lado, el propio artículo de partida sobre la clasificación de los ejemplares del yacimiento de Laño confirmaba un hecho que ya se conocía de manera previa y es que los dientes sin dentículos aparecen claramente diferenciados de los que sí que presentan esta característica morfológica; por este motivo, quizás resulte ineficaz utilizar una herramienta de análisis como el gráfico de dispersión bidimensional que se genera tras el «análisis de componentes principales» en el que —como se ha descrito— se van a visualizar las dos dimensiones que mejor diferencian a los elementos si ya se sabe de antemano que una de ellas va a estar ocupada por el número de dentículos. El motivo es que, si una de las dos variables descriptivas ya está ocupada, el gráfico sólo nos aportará información sobre la segunda.

Sin embargo, generar dos bases de datos separadas (una con los dientes que no tienen dentículos y otra con los que sí) y estudiarlas de manera separada (habida cuenta que las especies de uno y otro grupo no se solapan[20]) aporta dos ventajas:

1) En primer lugar, permite considerar el atributo «existencia de dentículos» cual, además, es de carácter categórico (Sí/No), lo cual era una de las intenciones iniciales que dieron origen a este proyecto[21].

2) En segundo lugar, una vez separadas las bases de datos para su estudio particularizado, este factor de presencia o no de dentículos que tanto afecta a los resultados de la clasificación conjunta se suprime por lo que queda más espacio para que otros factores (combinación de variables) se manifiesten. Lo cual resulta especialmente importante si

[20] Éste es un aspecto a discutir con los paleontólogos, en la medida en que se conozca especies que puedan presentar dientes con y sin dentículos… aunque, evidentemente, reconocer esta variabilidad puede llegar a suponer que varias de las premisas de las que se parte a la hora de establecer un método de clasificación automática deben ser replanteadas.

[21] En una consideración más detallada quizás se pueda pasar de una categorización dicotómica: «con dentículos» frente a «sin dentículos» a clasificaciones más amplias que puedan establecer diferentes clases en función de la forma de los dentículos (redondeados, triangulares, etc.), la presencia de los dentículos en ambos filos del diente (frente a los dientes con dentículos en un único filo), etc. Estos son aspectos que se apuntan como posibles vías de extensión del trabajo.

—como se hace en los gráficos de dispersión 2D— el número de factores que se retienen para su visualización está limitado.

A modo de ejemplo, se presenta a continuación el resultado del «análisis discriminante» de los dientes del yacimiento de Laño, como puede verse los dos grupos que se presentan a la derecha son los que corresponden a los dientes sin dentículos mientras que los de la izquierda son los que sí que tienen dentículos. Por lo tanto, el primer factor discriminante (de los dos disponibles en la representación gráfica) ya queda ocupado por esta característica.

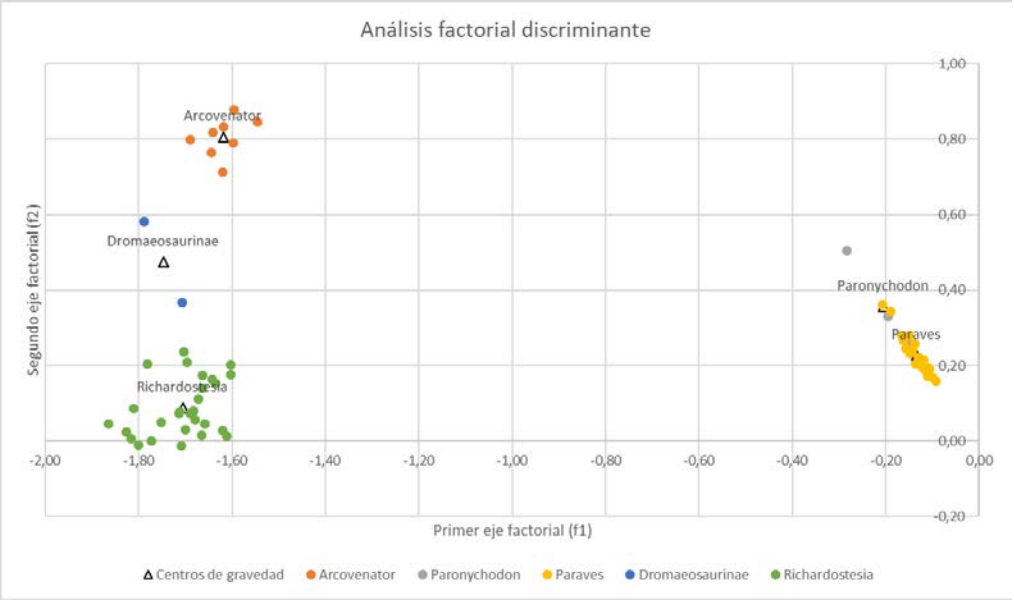


Fig. 21.- Representación de los dientes del yacimiento de Laño en los ejes factoriales del «análisis discriminante» considerando todas las clases y en el que se aprecia que el primer eje se utiliza para separar las clases con y sin dentículos.

Sin embargo, si aprovechamos esta representación para decidir que podemos aislar los dos conjuntos de datos (dientes con y sin dentículos) podemos hacer un nuevo «análisis discriminante» en el que sólo se consideren los dientes con dentículos. El resultado se muestra a continuación y, en él, se puede ver que ahora el primer eje factorial (el que aporta mayor poder de discriminación) corresponde aproximadamente al eje 2 de la representación anterior, quedando un nuevo segundo eje factorial que aportará información más específica sobre la disparidad de los datos.

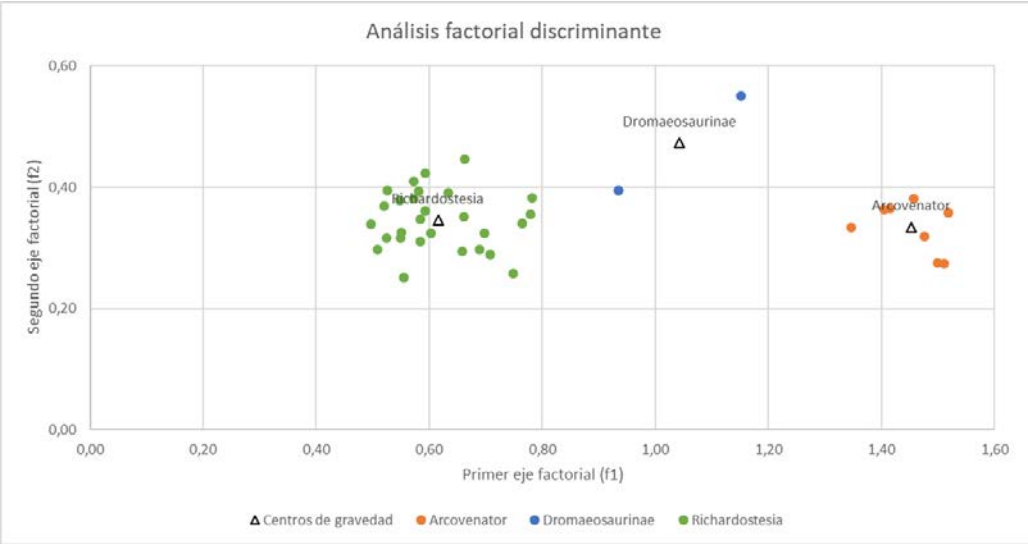


Fig. 22.- Representación de los dientes del yacimiento de Laño en los ejes factoriales del «análisis discriminante» considerando sólo las tres clases con dentículos y en el que se aprecia que el primer eje está especialmente condicionado por el tamaño, quedando el segundo eje factorial (vertical en este caso) como información más específica respecto al gráfico con todas las clases.

3) En tercer lugar, trabajar con bases de datos diferentes permite también seleccionar conjuntos de atributos particularizados para los elementos de cada base de datos, con lo que se pueden afinar mejor las clasificaciones dentro de cada subgrupo.

5.- Ejemplos de laboratorios de datos

La aplicación de las herramientas matemáticas para la clasificación de elementos ha tenido una amplia difusión en múltiples campos. Por este motivo, antes de definir una metodología de trabajo para el caso concreto de la paleontología, resulta interesante echar un vistazo a cómo se está abordando este mismo problema en otras áreas, con el fin de aprovechar la experiencia allí adquirida y poder detectar aspectos que quizás en los datos propios no resultan evidentes pero que sí que han sido considerados en esas otras áreas.

En concreto, se han recopilado dos ejemplos de tratamiento de datos:

1) L01.- Análisis de datos sedimentológicos[22], en el que se trata la determinación de las condiciones hidrodinámicas de un depósito a partir de la granulometría de los sedimentos.
2) L02.- Análisis de datos arqueométricos[23], referente a la determinación del origen de una producción cerámica a partir de la composición química de la pasta

[22] Información proporcionada por la profesora Ane Lopetegi.
[23] Información proporcionada por el profesor Javier Iñáñez.

De ambos se han redactado sendos documentos con la siguiente estructura:

a) Introducción general, indicando en qué contexto aparece la necesidad de clasificación y cuáles son los trabajos/autores que han definido los enfoques que se utilizan de manera común en este ámbito de la ciencia.
b) Objetivos, donde se expone qué es lo que se pretende determinar mediante el análisis.
c) Procesamiento de los datos paso a paso, se indica el proceso completo, desde la selección y toma de las muestras, su tratamiento antes de proceder a la medida de características, selección de atributos a medir, instrumental y métodos de medida utilizados, procesamiento de los datos, tipos de resultados que se generan (valores numéricos, tablas, gráficas, etc.) y forma de interpretarlos.
d) Aspectos de interés para el proyecto, partiendo de la información anterior se analiza qué ideas y métodos son aplicables al caso concreto de la clasificación de los dientes de dinosaurio.
e) Bibliografía, referencias generales sobre los métodos de clasificación aplicados en el área concreta que trata el documento (sedimentología o arqueometría).

Estos informes se incluyen como anexos en la versión impresa del informe y han servido para extraer ideas y tomar ejemplos de cómo se pueden aplicar las técnicas de clasificación. Sin embargo, al igual que las fichas relativas a las técnicas de clasificación, estos documentos sobre los «laboratorios de datos» se consideran documentación de trabajo por lo que no se incluyen en la versión disponible en el repositorio. En todo caso, sí que se considera de interés dejar reflejado lo indicado en el punto "d" de la lista anterior, es decir, los aspectos que, tras ver el procesamiento de los datos que se aplica en los diferentes ámbitos, se han considerado de interés para el caso concreto de la clasificación de los dientes de dinosaurio.

Tabla 1.- Aspectos aplicables de la revisión de los resúmenes sobre la aplicación de técnicas de clasificación en geomorfología y arqueometría.

Clasificación de sedimentos (geomorfología)	Un primer aspecto que resulta común al trabajo con la paleontología (en el caso concreto del artículo de los dientes que se ha visto como referencia inicial) es el uso de escalas modificadas respecto a los valores medidos en sus unidades originales. En este caso utilizan un logaritmo en base 2. El uso concreto de la base 2 con respecto a otras opciones (neperianos o base 10) posiblemente no tenga más trascendencia que las escalas en que se presentan los resultados (y su posible reflejo en los valores de referencia y umbrales que se establezcan), pero, en general, es de esperar que su empleo sea debido a la necesidad de realizar algún tipo de «normalización» sobre los datos originales. Por otro lado, es especialmente relevante la interpretación que se realiza de los parámetros estadísticos (asimetría, curtosis, etc.) en función de las características dinámicas de los procesos que causaron el depósito de los materiales. Este hecho es muy interesante porque permite dar sentido al tratamiento estadístico realizado.

Si comparamos el proceso descrito al que está empleado en la clasificación de los dientes en el yacimiento de Laño (Isasmendi et al., 2022) se pueden sacar algunos elementos que resulta interesante analizar:

a) Un primer aspecto que resulta común es el uso de un banco de datos de referencia (de cerámicas en este caso), que resulta análogo al que se utiliza de dientes en la aplicación a paleontología.

b) En segundo lugar, resulta curioso notar que en el caso de la arqueometría se recurre a una normalización mediante logaritmos, considerando que este proceso corrige variaciones debidas a las diferentes temperaturas de cocción de las piezas. De hecho, también se normaliza a través del logaritmo en el ejemplo de caso del que se parte en el yacimiento de Laño… si bien, el significado físico de esta normalización no tiene nada que ver con la que justifica su empleo en arqueometría.

c) En el caso visto de la arqueometría, los datos brutos obtenidos de la instrumentación tienen un proceso de ajuste mediante curvas de calibración, lo que supone una fase de preprocesado (aparte de la toma de logaritmos).

d) El dendrograma (como resultado) y la forma de obtenerlo son los que se ha analizado en la ficha del «análisis de conglomerados», es interesante notar que en esta aplicación no tienen por qué conocer de antemano el número de grupos que se van a formar y que, de hecho, uno de los fines de esta herramienta es determinar piezas que no pertenecen a los grupos anteriores (es decir, establecer posibles nuevas clases). En el caso del estudio paleontológico concreto que se ha mencionado no se emplea este método.

e) Sin embargo, la técnica que sí que es común es el «análisis de componentes principales» que en ambos casos se emplea para reducir la dimensionalidad del conjunto de datos de caracterización de los elementos. Como se puede ver, en el caso de los datos de espectrometría, la cantidad de variables (que corresponde a los contenidos de los diferentes elementos químicos) resulta muy elevada.

f) Al respecto de las variables empleadas, es interesante resaltar en lo que indican de las contaminaciones en las pastas cerámicas. Este fenómeno implica que las variables a utilizar para caracterizar de manera conveniente las cerámicas pueden no ser siempre las mismas ya que, en algunos casos, determinadas variables pueden no ser significativas e incluso resultar engañosas. Éste es un escenario que no se ha tenido en cuenta en el análisis paleontológico pero que quizás merezca atención.

g) Como forma de aplicación práctica, es destacable la preparación de un paquete propio en R y su difusión pública (a través de GitHub) para el uso de la comunidad científica y permitir la trazabilidad de los cálculos realizados.

6.- Diseño e implementación del procedimiento para el tratamiento de la base de datos

Con la información recopilada sobre la forma en que se está procesando la información paleontológica y las ideas traídas de los ejemplos de arqueometría y geomorfología, el siguiente paso consiste en confeccionar una metodología de procesado mejorada para los datos de los dientes. Esta metodología consta de dos partes: por un lado, se debe determinar la secuencia de operaciones a realizar (es a lo que se dedicará el punto siguiente de este documento); por otro lado, también resulta necesario definir la forma en que ésta deberá llevarse a cabo (que es sobre lo que se hablará en el presente apartado).

La meta no consiste en diseñar una herramienta que, de manera automática y como una caja negra, sea capaz de clasificar los dientes de dinosaurio. En efecto, ya que se ha visto que existe una gran diversidad de matices y consideraciones que han de tenerse en cuenta y, de hecho, puede llegar a ser discutible incluso el hecho de que dicha herramienta sea factible. El problema de la clasificación de dientes está abierto a debates y nuevas incorporaciones, por lo tanto, no existe un conocimiento establecido y cerrado que pueda plasmarse en un conjunto de reglas y algoritmos que den como resultado una asignación a una especie determinada que se pueda considerar inequívoca.

Por el contrario, lo que sí que resultaría útil es disponer de una herramienta interactiva que permita ir incorporando fases del procesamiento y que, mediante secuencias de prueba y error permitan articular un debate, ir generando hipótesis y consolidando conocimiento sobre el tema. Para ello, es necesario que la herramienta sea intuitiva y sencilla de implementar y que pueda usarse de manera compartida por los diferentes tipos de especialistas de forma que sirva como medio de comunicación (en concreto, entre la parte «paleontológica» que conoce los datos de partida y puede dar sentido a los productos en el contexto real y la «matemática» que entiende en qué medida las operaciones con los datos son adecuadas a las características de la información disponible y cuál es el nivel de significación de los resultados que se van obteniendo).

Por lo tanto, lo que se está buscando es un entorno de trabajo y, a este respecto, se estuvieron valorando varias opciones y se optó por un sistema gráfico denominado Orange[24], que es un software de minería de datos —desarrollado por la universidad de Liubliana— que trabaja utilizando una interfaz gráfica mediante la conexión de componentes.

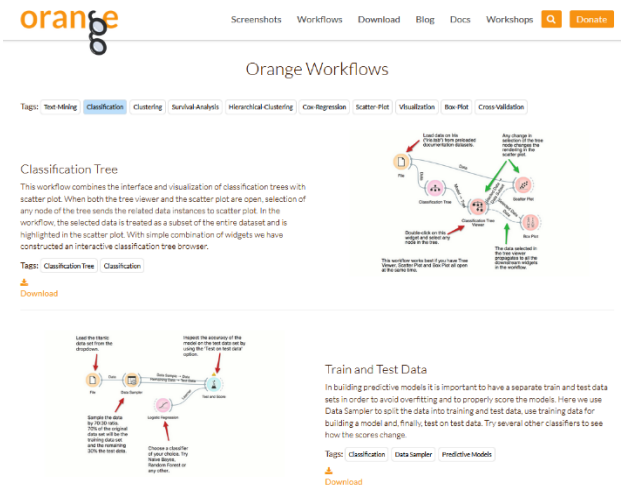


Fig. 23.- Ejemplos explicados de flujos de trabajo creados con el software Orange.

[24] https://orangedatamining.com/

El programa es gratuito y su manejo resulta muy intuitivo por lo que resulta interesante para plantear procesos de procesado de información de manera bastante rápido. Asimismo, dispone de un variado surtido de herramientas matemáticas por lo que el procesado resulta bastante potente. La revisión preliminar de este software y su posible adecuación al problema de los dientes se plasmó en un nuevo documento[25]:

- L03.- Procesamiento gráfico interactivo (con `Orange`).

El cual también se presenta como anexo a la versión impresa del proyecto y del que extraemos aquí algunos puntos de interés.

Como se ha avanzado, la forma de expresar un flujo de trabajo se hace situando sobre un espacio de trabajo diferentes tipos de componentes (ficheros, filtros, operaciones, salidas gráficas…) que se van conectando y ajustando los parámetros para indicar las características de las conexiones y las operaciones que se tienen que realizar.

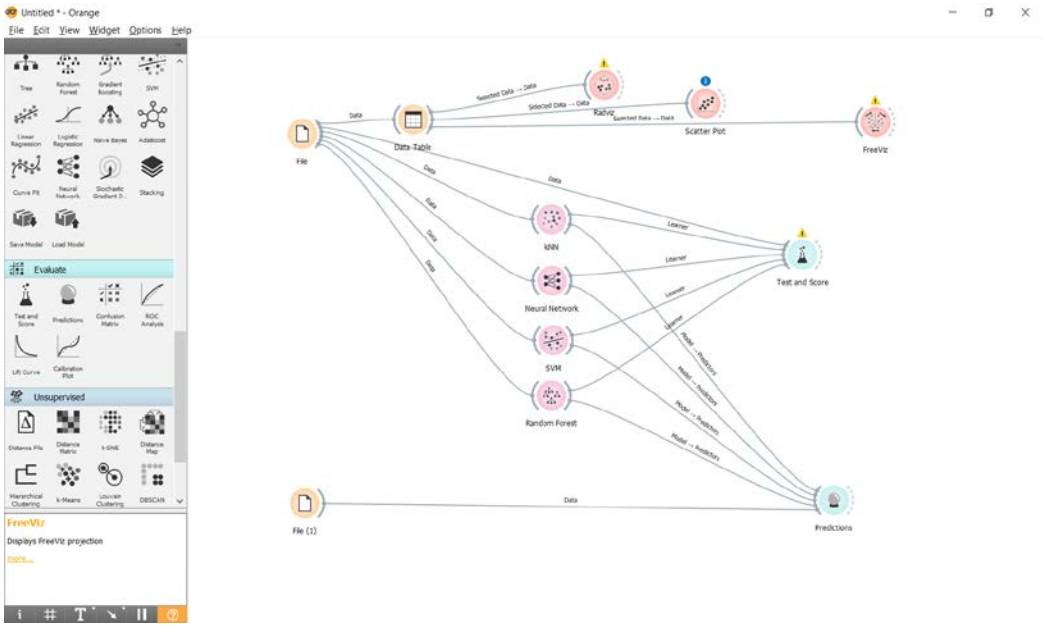


Fig. 24.- Vista de un ejemplo de flujo de trabajo para el procesado de los datos.

Una vez establecidas las conexiones, los resultados se van actualizando permanentemente; por ejemplo, si se realiza una visualización de una selección de elementos y se cambia la selección realizada o si en el control relativo a un fichero de entrada de datos se selecciona un nuevo fichero con nuevos datos. Esto permite separar la definición del proceso de trabajo de los propios datos y, asimismo, facilita la reutilización del mismo flujo de trabajo en repetidos conjuntos de datos.

Se ha analizado la información de formación sobre el programa (que está compuesta de varias colecciones de vídeos cortos en los que se van mostrando los diferentes controles y su uso en ejemplos desarrollados) y se han ido aplicando a casos con las bases de datos de los dientes que están disponibles. En concreto, se ha trabajado con la carga y preprocesamiento de datos, la representación gráfica utilizando diferentes tipos de visualizaciones, las herramientas de

[25] Información proporcionada por el profesor Beñat García.

clasificación no supervisada (que corresponden a la técnica de los «análisis de conglomerados») y de reducción de la dimensionalidad («análisis de componentes principales») así como diferentes controles para diversos tipos de «análisis discriminantes».

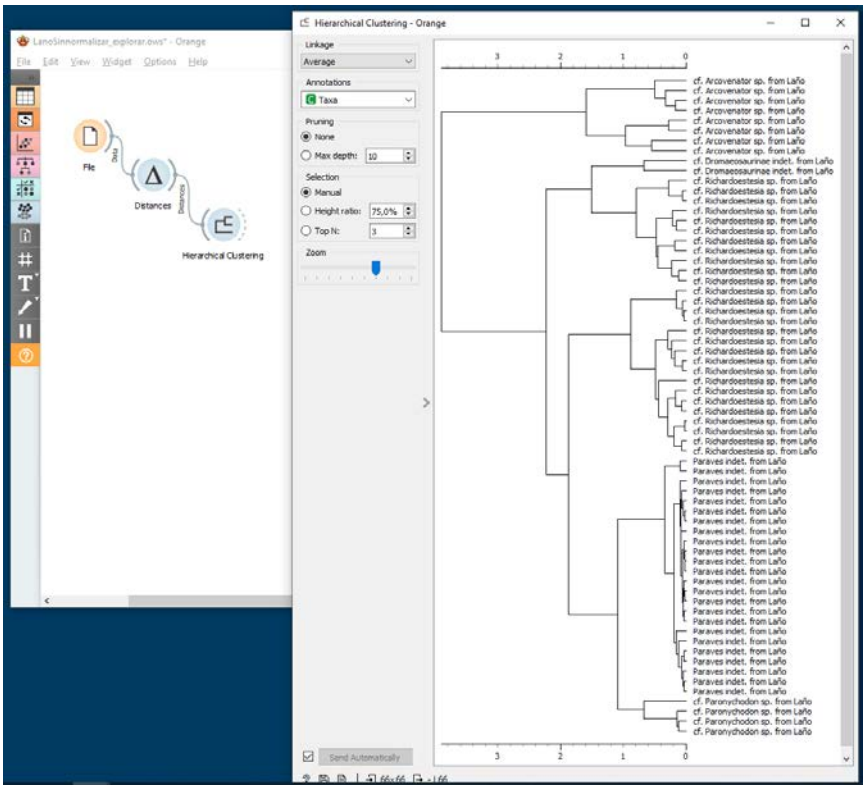


Fig. 25.- Ejemplo de uso del control del «análisis de conglomerados» (denominado como "*hierarchical clustering*" en el programa) aplicado a la base de datos del yacimiento de Laño.

También se han examinado los controles de algunas herramientas que no se habían llegado a analizar en las fichas de las técnicas seleccionadas como, por ejemplo, los «árboles de decisión» o la «regresión logística». El hecho de que exista una amplia variedad de técnicas adicionales, resulta muy interesante ya que da posibilidades de extender el repertorio de herramientas de análisis progresivamente (conforme se vayan estudiando cada una de ellas con el fin entenderlas debidamente y controlar cómo se deben preparar los datos de entrada y cómo se deben interpretar las salidas generadas).

7.- Diseño e implementación del procedimiento para el tratamiento de la base de datos

La base de datos de Laño dispone de 228 elementos o especímenes, 23 variables cuantitativas y 1 cualitativa. Podemos sentir la tentación de usarla completa, pero eso puede llevarnos a incluir en la metodología datos indeseados que pueden falsear los resultados y las conclusiones que podemos sacar y/o tratar como iguales elementos claramente distinguibles.

Por ello es importante "limpiar" nuestra base de datos y fragmentarla en sub-bases para poder extraer todo el potencial de las técnicas analíticas.

7.1.- Cribados previos

1) Primer cribado: creación de sub-bases de datos

El primer cribado consiste en una toma de decisión basada en el conocimiento. A veces tratamos de trabajar con bases de datos lo más extensas posibles por aquello de que a mayor tamaño de muestra más certeras son las conclusiones a las que se llegan. Sin embargo, a menudo, esto puede llevarnos a perder información si estamos tratando con datos que, de antemano, ya sabemos que pertenecen a grupos o conjuntos diferentes.

Nuestro conocimiento previo sobre el área de trabajo nos puede llevar a identificar de forma clara conjuntos de especímenes dentro de una misma muestra o base de datos teniendo en cuenta el valor que toma alguna de las variables. En tal caso, lo conveniente sería utilizar dicha variable (que puede ser cuantitativa o cualitativa) para dividir nuestra base de datos en diferentes sub-bases de datos. De esta manera, los elementos de la muestra dentro de la misma sub-base tendrán más en común que con los elementos de las otras sub-bases. La ventaja principal de crear sub-bases de datos es que podremos definir técnicas analíticas adecuadas para cada una de las sub-bases que resalten más claramente sus diferencias (separación en la representación gráfica), es decir, podremos diseñar análisis optimizados al conjunto contenido en cada sub-base.

En nuestra base de datos de Laño la variable cualitativa disponible es la de *presencia de dentículos* que puede tomar valores "sí" o "no" y que queda reflejada en las dos variables seleccionadas MDD y DDD, ya que si la variable "presencia de dentículos" toma el valor "no", entonces sus valores en las variables MDD y DDD serán 0 en ambas. Por lo tanto, fijándonos en los valores de las variables seleccionadas MDD y DDD que toman los especímenes de la base sólida, podremos distinguir dos grupos: los que tienen MDD=DDD=0 y los que no (estos últimos podrán tener valores distintos de cero en ambas variables o únicamente en una de ellas). Así formaremos dos sub-bases:

$$\rightarrow sub-base\ A\text{: }especímenes\ que\ no\ presentan\ dentículos \begin{cases} - grupo\ Paranychodon \\ - grupo\ Paraves \end{cases}$$

$$\rightarrow sub-base\ B\text{: }espeecímenes\ que\ sí\ presentan\ dentículos \begin{cases} - Arcovenator \\ - Dromaeosaurinae \\ - Richardoestesia \end{cases}$$

Dependiendo de la base de datos disponible o de lo que se desee profundizar en la clasificación, también podría ser interesarte añadir un posterior cribado que influya dentro de cada sub - base: por ejemplo, según el tamaño del diente, ya que podría ser interesante distinguir entre dientes pequeños y grandes dentro de los mismos grupos y ver si gráficamente están igual posicionados que sus homólogos grandes.

2) Segundo Cribado: selección de variables a considerar

Antes de comenzar con el análisis estadístico de los datos de cada sub-base, es importante identificar qué variables son menos relevantes por aportar información que ya queda reflejada en otras variables o por provenir de la combinación de otras variables. De esta forma en lugar de tener que trabajar con un gran número de variables, podremos centrar nuestro análisis en aquellas más representativas.

En la base de datos completa de Laño aparecen 23 variables cuantitativas y una cualitativa. Sin embargo, algunas de ellas aportan información que, desde el punto de vista del análisis estadístico, es redundante. Un caso, por ejemplo, es la variable CBR (*crown base ratio*) está definida como la división CBW/CBL (*crown base width / crown base length*) y que podríamos excluir de los posteriores análisis si incluimos las dos que la definen y la normalización mediante logaritmos. Otro caso sería la variable que mide la densidad de dentículos por mm y la que la mide por cada 5 mm, desde el punto de vista estadístico con coger una de ellas (se puede deducir una de otra) sería suficiente. Empleando este criterio de selección de variables se considera adecuado hacer uso de las siguientes 5 variables cuantitativas para la sub-base B:

$$\rightarrow \ tamaño\ del\ diente \begin{cases} CBL: altura\ (crown\ height) \\ CBW: longitud\ de\ la\ base\ (crown\ base\ length) \\ CH: grosor\ de\ la\ base\ (crown\ base\ width) \end{cases}$$

$$\rightarrow \ presencia\ de\ dentículos \begin{cases} MDD: densidad\ en\ cara\ mesial\ (mesial\ denticle\ mesial \\ DDD: densidad\ en\ cara\ distal\ (distal\ density\ denticle) \end{cases}$$

Para la sub-base A, sin embargo, al tratarse de especímenes que carecen de dentículos consideraríamos las 3 variables relacionadas con el tamaño del diente.

3) Tercer cribado: selección de elementos válidos de la muestra

Finalmente se debe tener en cuenta que tratar con variables con valor "vacío" (ausencia de valor) en algún/algunos elemento(s) puede afectar a la posterior identificación del espécimen por causa de una incorrecta representación gráfica. Esto se debe a que, dependiendo de la técnica analítica empleada, a los valores "vacíos" se les adjudica matemáticamente un valor de cero o de 1 (si se considera la variable x+1) cuando este valor adjudicado no es representativo ya que la realidad es que no se dispone de él.

Por ello, tras seleccionar las variables con las que se va a trabajar y si disponemos de sub-bases de datos extensas, el tercer cribado consiste en retirar de dichas sub-bases aquellos especímenes que estén incompletos y no se hayan podido completar virtualmente. De este modo tendremos unas sub-bases robustas, formadas por elementos con valores no vacíos en todas las variables seleccionadas, a partir de las cuales podremos clasificar de forma fiable nuevos especímenes.

En la muestra de Laño, de un total de 228 dientes, se seleccionan 28 especímenes para la sub-base A que contedrá dos tipos de dinosaurios (4 Paranychodon y 24 Paraves) y 38 especímenes para la sub-base B con tres tipos de dinosaurios (8 Arcovenator, 2 Dromaeosaurinae y 28 Richardostesia).

Una vez finalizados los tres cribados, que pueden ser recurrentes para seguir optimizando las sub-bases, procederíamos a la aplicación de la técnica analítica más adecuada para cada una de ellas.

7.2.- Metodología de análisis para la sub-base de datos sin dentículos (sub-base A)

En el estudio publicado, los dos grupos de dinosaurios presentes en esta sub-base A aparecen solapados en la gráfica del análisis de componentes principales. En dicho estudio también se presenta un análisis discriminante donde no se logra diferenciar los dos grupos. Es importante indicar que para esta sub-base A únicamente se trabaja con las tres variables relacionadas con el tamaño del diente por lo que cobra sentido plantear un diagrama de dispersión en 3D que abarque el espacio vectorial formado por las tres variables:

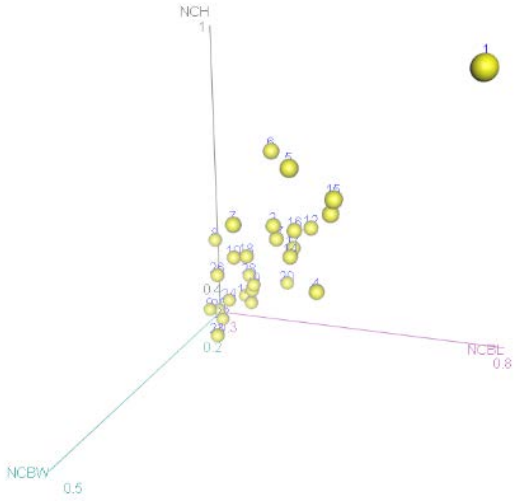


Fig. 26.- Representación gráfica de tres variables factoriales de la sub-base A.

Los elementos de los dos posibles grupos de dinosaurios se solapan y no es posible encontrar ninguna característica propia para cada uno de los grupos en base a la información de estas variables.

Esto no ocurre con los datos del sub-grupo B tal y como se observa en el siguiente diagrama de dispersión en 3D donde se han empleado dos variables relacionadas con el tamaño del diente y una tercera relacionada con la presencia de dentículos para la base de datos completa:

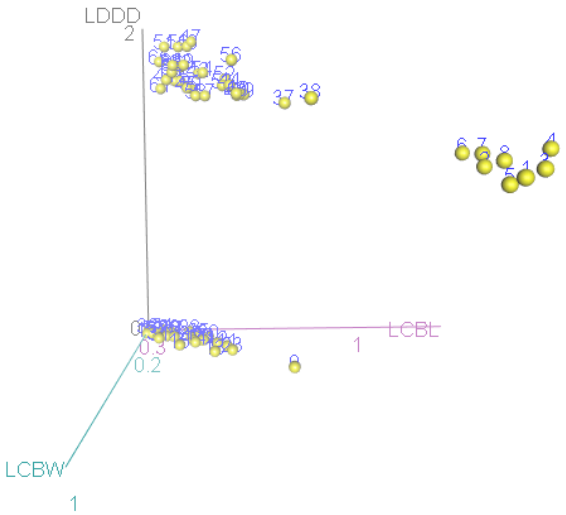


Fig. 27.- Representación gráfica de tres variables factoriales de la sub-base B.

En este caso los elementos de las sub-base A quedan en el plano inferior amontonados mientras que los elementos de la sub-base B quedan a una cota superior y parecen describir 3 agrupamientos.

7.3.- Metodología de análisis para la sub-base de datos con dentículos (sub-base B)

En el estudio publicado, los tres grupos de dinosaurios presentes en esta sub-base B aparecen solapados en la gráfica del análisis de componentes principales. En dicho estudio también se presenta un análisis discriminante que logra reorganizar los datos y eliminar la "mezcla" de los puntos de diferentes grupos reduciéndolos en un simple solapamiento a lo largo del eje vertical. Tras diversos análisis, se ha obtenido una separación (discriminación) total de los grupos a lo largo de los dos ejes mediante un procedimiento combinado:

Se aplica un análisis de componentes principales con las 5 variables seleccionadas para esta sub-base y posteriormente un análisis discriminante. A continuación, se recoge la gráfica tras estos dos análisis combinados:

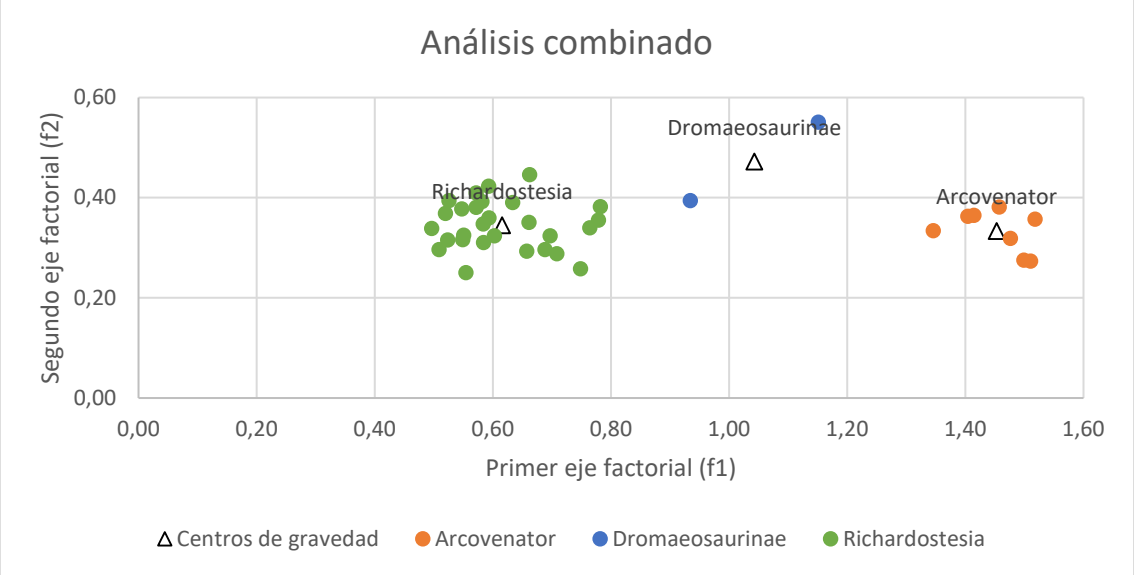


Fig. 28.- Análisis discriminante aplicado a la sub-base B utilizando los cinco atributos disponibles.

Esta representación gráfica es la que se obtiene directamente de la aplicación combinada de las dos técnicas analíticas mencionadas: los grupos pertenecientes a cada tipo de dinosaurio quedan discriminados en el primer eje factorial f1. De esta forma, un nuevo espécimen sería adjudicado a uno y otro grupo de dinosaurios teniendo en cuenta, exclusivamente, su valor de f1, ya que en el segundo eje factorial f2, los grupos no son distinguibles. Por ello, una vez aplicada la técnica analítica combinada, puede resultar interesante hacer un cambio de ejes. En este caso, un giro de ejes nos permitirá. Para esta sub-base B, un giro a unos nuevos ejes ortogonales f1+f2 y f1-f2 nos permite discriminar los grupos en ambos ejes y tener, de esta manera, dos criterios para adjudicar un nuevo espécimen: su valor en el eje f1+f2 y también su valor en el eje f1-f2.

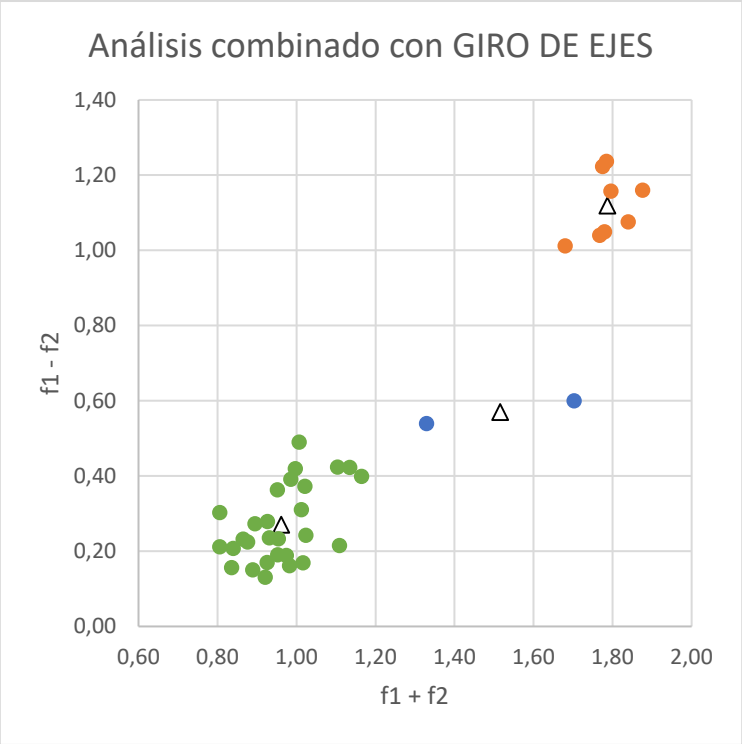


Fig. 29.- Aplicación de un cambio de ejes adicional para aumentar la separación entre las clases al queden separadas en los dos ejes resultantes tras este último cambio.

7.- Resultados y discusión

Respecto al proceso de trabajo que se propone, éste supone prescindir del enfoque en el que las técnicas de clasificación funcionan como cajas negras y reemplazarlo por un proceso de diálogo que vaya profundizando en el análisis de los datos de manera progresiva y en el que vayan combinándose preguntas y respuestas que atiendan de manera conjunta aspectos relativos a la paleontología y a la matemática. Incidiendo en esta línea, lo que se pretende recalcar es que es la propia generación del proceso más adecuado para el procesamiento de los datos —el cual se realiza a través de la comunicación y el ciclo de prueba y error— lo que realmente genera conocimiento y es la parte que debe valorarse en mayor medida (al menos, de manera preferente a disponer de un esquema de procesamiento preestablecido cuyo fundamento resulte desconocido).

La implementación del esquema propuesto mediante un lenguaje de programación concreto sistematiza el trabajo de tener que ejecutar cada parte del diagrama de flujo de manera individual. Este lenguaje podría ser "R" (tal como se hizo en el caso del laboratorio de datos de cerámicas para el ejemplo analizado de arqueometría), aunque, para nuestro caso, se ha explorado la opción de utilizar un lenguaje visual (`Orange`).

El esquema gráfico que se ha utilizado pretende facilitar la comunicación entre los diferentes perfiles de profesionales que deben intervenir en el proceso de análisis y clasificación de los dientes. Perfiles que, en general, no tienen por qué disponer de un conocimiento extensivo de la notación y fundamentos de las herramientas matemáticas, pero sí que deben entender su fundamento, conexión y las características tanto de la información de partida como de los resultados que se generan.

Este enfoque es adecuado para una aproximación progresiva al problema de la clasificación (y, en general, extensible a cualquier otro problema de aplicación de técnicas matemáticas), ya que permite comenzar con un conjunto reducido de técnicas y utilizando cada una de ellas a un nivel básico. Aun así, ya desde el inicio, se pueden crear flujos de trabajo que sean funcionales y permitan obtener resultados y conclusiones, lo cual permite generar dinámicas de trabajo positivas.

En la medida en que se disponga de más tiempo y capacidad de ir mejorando el sistema, éste se puede ampliar incorporando nuevas técnicas al conjunto de las disponibles y/o refinando la forma en que se utiliza cada técnica de forma que se permita trabajar con una mayor variedad de datos de partida y se obtengan nuevos tipos de resultados.

Por otro lado, no se debe desdeñar el valor que tienen los diagramas de flujo en la propia conceptualización del proceso de trabajo, sirviendo como elementos de diseño, comprensión y comunicación del trabajo realizado.

Sería apropiado realizar un estudio sobre el número de variables (medidas) que entran en los diferentes análisis de forma que se eliminen aquellas que sólo aportan ruido (y, por consiguiente, entorpecen el discernimiento de los grupos, además de involucrar una mayor carga de cálculo), para ello se debería proseguir analizando la aportación individual de cada variable a los ejes principales definidos en los diferentes análisis así como realizar estudios de incorporación/eliminación progresivos que permitan ir viendo las situaciones que se producen en diferentes escenarios.

El tema de la «normalización» de las variables que se utilizan en el «análisis de componentes principales» es un tema abierto y que merece una reflexión detallada que conjugue de manera adecuada tanto la corrección del cálculo, su interpretabilidad en el ámbito matemático y su representatividad en el problema morfológico.

Se comentó que uno de los problemas que pueden existir al intentar comparar tablas de datos morfométricos son los diferentes criterios que se hayan utilizado para la selección de las medidas características y su determinación (por ejemplo, entre qué dos puntos se debe medir una altura). El hecho de proporcionar los modelos 3D completos permite paliar este problema ya que cualquier nuevo investigador puede visualizarlos y extraer a partir de ellos las medidas que considere convenientes (si no existen previamente en las tablas descriptivas del espécimen) así como comprobar si las medidas recogidas corresponden o no al mismo criterio de determinación que espera utilizar en el nuevo análisis. En todo caso, es interesante constatar la fuerte relación que existe entre esta línea de investigación (el análisis de las herramientas matemáticas) y la relativa a la documentación 3D de los especímenes ya que esta última es la que aporta los datos de partida de los análisis, determinando la precisión de partida.

También es interesante notar que los valores de las tablas no suelen venir acompañados de las incertidumbres asociadas a las medidas y que estas incertidumbres no se consideran posteriormente ni en el cálculo ni en el análisis de los resultados.

La selección de la base de datos de referencia para las comparaciones, si se trata de dientes compatibles con la época y la zona geográfica o se ha ampliado el marco (con los riesgos de caer en malas interpretaciones que ello puede suponer).

Sobre el uso de los análisis estadísticos como prueba para considerar que dos elementos pertenecen o no a una misma categoría. Existe una fuerte dependencia de si los atributos

considerados en el análisis son o no significativos de esta diferencia, de lo contrario no existirá relación entre lo que se muestre en el análisis matemático y la realidad. Por este motivo resulta necesario considerar todas las variables que sean determinantes y excluir las que no lo sean.

También la evaluación en porcentajes de clasificación correcta no tiene en cuenta la similaridad de las clases y la diferente importancia que tienen unas asignaciones erróneas frente a otras. Quizás sería interesante llegar a determinar algún tipo de medida ponderada que redujese la importancia de los errores cuando estos son justificables y les diera más importancia cuando indiquen problemas reales del algoritmo de clasificación empleado.

El estudio es extensible en el futuro, considerando nuevas técnicas de clasificación o tratamiento de datos que se puedan ir integrando en el esquema gráfico, así como profundizando en el potencial de cada una de las técnicas, extrayendo, por ejemplo, una información más rica sobre la capacidad discriminante de los factores conservados, utilizando los resultados para seleccionar mejor las variables de partida utilizadas en los análisis (a través del estudio de la contribución de cada una a las soluciones) o analizando la posibilidad de aumentar el número de factores discriminantes (por ejemplo, pasando de las representaciones gráficas bidimensionales de los ejes factoriales a visualizaciones 3D interactivas), entre otras posibilidades.

Por otro lado, resultaría interesante profundizar en el significado de los ejes factoriales que resultan de los diferentes análisis (en particular los que son establecidos a través del «análisis de componentes principales»). Esta parte también requiere del análisis conjunto de la paleontología (en la parte del significado) y de la matemática (en lo referente a cuantificar la aportación de las variables iniciales). Esta información permitirá afinar mejor las caracterizaciones que se establecen (conjunto de atributos que se miden de los diferentes especímenes), enfocándose en aquellos que efectivamente aportan poder discriminante y descartando aquellos que no contribuyen a esta tarea[26].

En una situación ideal los patrones identificados en la clasificación automática (o asistida) deberían ser coherentes con el conocimiento que se dispone sobre los especímenes y la taxonomía de clasificación por lo que, en gran medida, el desarrollo de ambas ramas debe discurrir en paralelo.

El uso de los diagramas de dispersión bidimensionales es un recurso habitual que está ligado a las posibilidades de análisis (y representación) visual soportadas por el papel. En un contexto más actual soportado por la pantalla del ordenador, también es posible pensar en incorporar una tercera dimensión (es decir, un tercer eje factorial) en representaciones interactivas 3D.

[26] En todo caso, el hecho de que una determinada característica no resulte útil en la clasificación no implica que no pueda ser interesante para otras aplicaciones y, por lo tanto, merezca ser recogida; lo que sí que permite establecer es que quizás no deba emplearse en el cálculo de la clasificación.

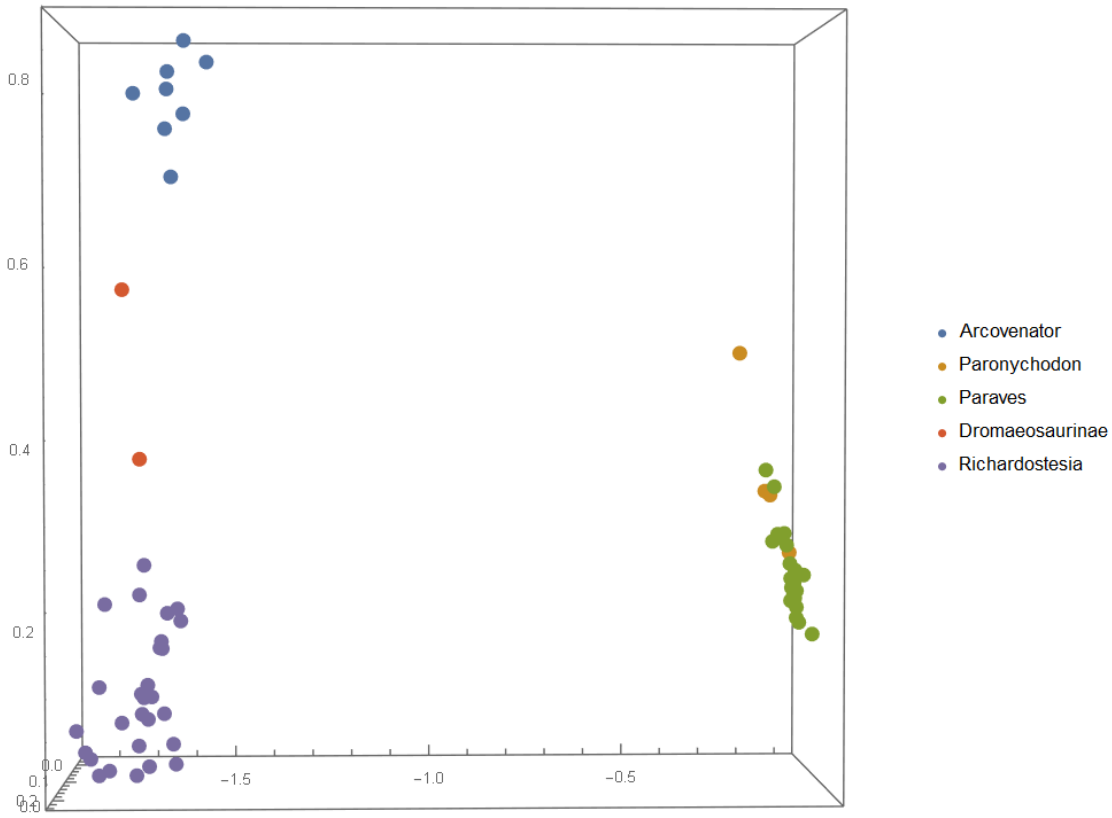


Fig. 30.- Vista superior del cubo con la representación 3D de los tres primeros ejes factoriales resultantes del «análisis discriminante» de los dientes del yacimiento de Laño.

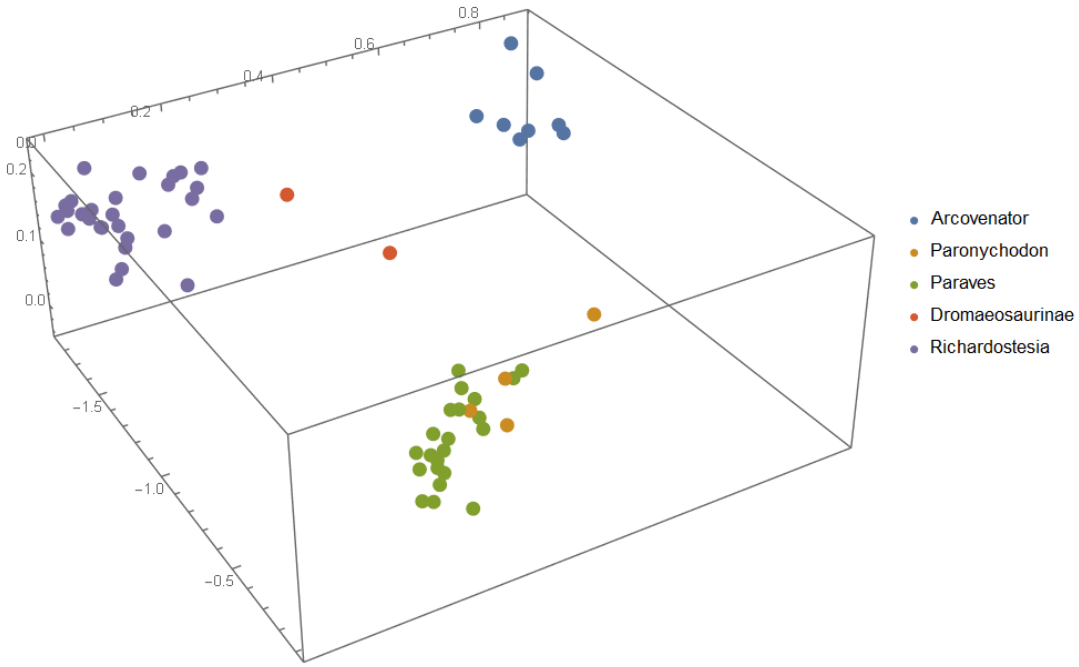


Fig. 31.- Vista perspectiva del cubo con la representación 3D de los tres primeros ejes factoriales resultantes del «análisis discriminante» de los dientes del yacimiento de Laño.

Finalmente, al respecto del programa de modelado gráfico (Orange), se trata de una herramienta muy potente, sencilla e intuitiva de utilizar. Permite crear procesos de trabajo que se muestran gráficamente por lo que son fácilmente interpretables y que pueden reutilizarse con diferentes bases de datos. Además, una vez que se dispone del proceso se pueden ir analizando los resultados, haciendo cambios en los parámetros, seleccionando conjuntos de datos, etc. convirtiéndose en una forma de análisis visual de la información que permite descubrir patrones y plantear hipótesis sobre los registros de la base de datos. En definitiva, se considera que es una herramienta de gran interés en el contexto del presente proyecto.

Lo que queda ahora es la fase de transferencia e implementación que consistirá en juntarse con los paleontólogos y, de manera interactiva, ir diseñando el flujo de componentes sobre la interfaz gráfica para ir viendo e interpretando los resultados en un entorno de trabajo colaborativo en el que se establezca un diálogo que aúne los conocimientos sobre paleontología y las técnicas de análisis matemático.