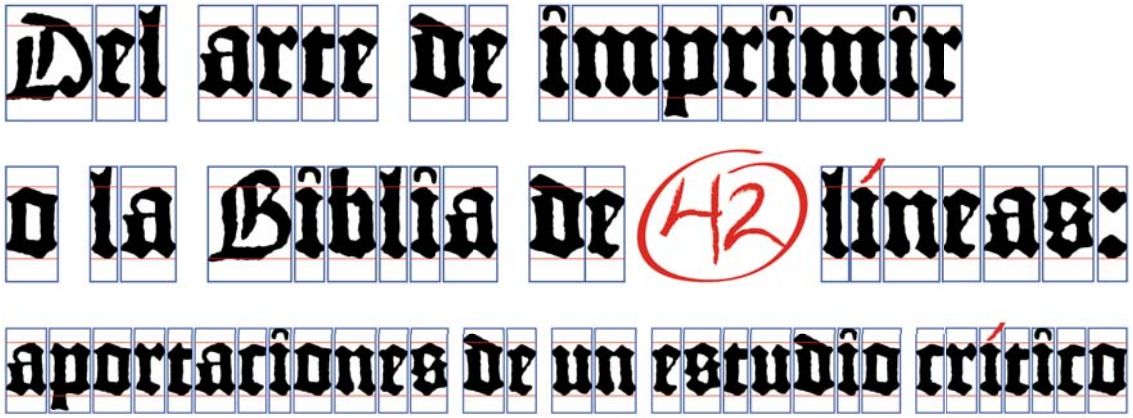
Del Arte de imprimir o la Biblia de 42 líneas: aportaciones de un estudio crítico

Luz María Rangel Alanís



Tesis doctoral presentada por:

Luz María Rangel Alanís

Dirigida por:

Dr. Enric Tormo Ballester
Universidad de Barcelona

Dr. Aureli Alabert Romero
Universidad Autónoma de Barcelona

Programa de doctorado:

Las revoluciones tipográficas
Bienio 2000-2002

Universidad de Barcelona
Facultad de Bellas Artes
Departamento de Diseño e Imagen

Abril, 2011.

Capítulo 7

Secundum gratiam Dei, quae data est mihi,
ut sapiens architectus fundamentum posui:
alius autem superaedificat.
Unusquisque autem videat quomodo superaedificet.

1 Ad Corinthios 3:10

. . .

Conforme a la gracia de Dios que me ha sido dada,
yo como perito arquitecto puse el fundamento,
y otro edifica encima;
pero cada uno mire cómo sobreedifica.

1 Corintios 3:10

7. Clasificación de tipos en la Biblia de 42 líneas

7.1 Introducción

Se pretende, a partir de letras impresas con distintos *tipos*, y a la vista de las formas impresas de dichas letras, decidir si se puede afirmar que los tipos correspondientes no proceden todos de la misma *matriz*.

La letra impresa tiene unos "errores" (o "desviaciones") respecto a la forma ideal derivada de un tipo, debidos a la tinta, al entintado, a la humedad, al papel, etc. También cada tipo metálico derivado de la misma matriz puede tener desviaciones respecto a la forma ideal dada por la matriz (errores de origen o debidos a desgaste por el uso). Estos errores pueden considerarse aleatorios, es decir, no sistemáticos. Nos interesa "filtrar" estas desviaciones para observar si hay una "diferencia estructural" entre letras, que pueda ser debida a que proceden de matrices diferentes.

Naturalmente, letras impresas con el mismo tipo proceden de la misma matriz y se parecerán mucho entre sí. Queremos precisamente estudiar un conjunto de letras que hayan sido impresas todas ellas con tipos distintos. De este modo, si se observan grupos de letras que se parecen entre sí dentro de cada grupo, y que son diferentes de las de los otros grupos, sabremos que estas diferencias sólo pueden ser debidas a una procedencia matricial distinta.

Para garantizar que los tipos son todos distintos entre sí, no hay más que tomar letras impresas en una misma página. Se ha elegido la primera página del Evangelio de Mateo, y en ella se han estudiado 10 letras i y 21 letras a. Las i se tomaron de la columna I alternando un renglón si y otro no empezando desde el primer renglón es decir líneas impares; en el caso de las a es lo mismo pero se inicia en el segundo renglón por lo que las líneas seleccionadas son las pares. Por un lado, el número de 10 es suficientemente pequeño para hacer una clasificación mediante el llamado *cluster analysis* ("análisis de conglomerados") con el que se puede obtener el *clustering* (agrupación) óptimo bajo un criterio prefijado. Por otra parte, el número de 21 letras a permite más posibilidades de descubrir pautas de clasificación, aunque se debe recurrir a un método aproximado de cluster analysis debido a la enorme cantidad de tiempo de cálculo que requeriría intentar obtener el clustering óptimo.

En el Apartado 7.2 se introduce la metodología del cluster analysis. El Apartado 7.3 está dedicado a la descripción detallada del procedimiento que se ha seguido para medir efectivamente las *distancias* entre pares de letras impresas, que son los datos iniciales necesarios para aplicar el cluster analysis. El Apartado 7.4 explica cómo se realiza un cluster analysis exacto, a la vez que se obtienen los resultados

para las diez letras i de la primera página del Evangelio de Mateo. El cluster analysis aproximado se describe en el Apartado 7.5, y se ilustra otra vez con las letras i, para comparar los resultados con los métodos exactos; esta comparación nos permitirá asegurar que el estudio mediante el método aproximado de las veintiuna letras a, que se realiza en el Apartado 7.6 tiene en efecto sentido. La aplicación en si del método aproximado, llamado *hierarchical clustering* (clustering jerárquico), se realiza en el Apartado 7.6.1, mientras que en el Apartado 7.6.2 se aplica un método estadístico de detección de datos anormales para validar los clusters obtenidos en el primero.

No es posible representar gráficamente en un plano cada letra en una posición, de modo que las distancias en el plano sean exactamente las distancias medidas entre las letras. No obstante, sí se puede hacer una representación aproximada, en dos y en tres dimensiones, lo que nos permitirá corroborar gráficamente las conclusiones numéricas a las que llegaremos. A esta representación aproximada, llamada *multidimensional scaling*, se dedica el Apartado 7.7.

Las letras a, ofrecen la posibilidad adicional de estudiar sus contornos interiores, lo que permite comprender mejor no sólo la construcción de matrices sino también la de punzones. En el Apartado 7.8 se han clasificado, con la misma metodología que para las letras completas, los contornos interiores inferiores de las mismas veintiuna a, y se obtiene una tabla cruzada de ambas clasificaciones.

Nótese que si se usan letras impresas procedentes de páginas distintas, que pueden por tanto proceder de un mismo tipo, deberían poder observarse tres niveles de similitud entre letras impresas: Letras muy parecidas entre sí, apareciendo obligatoriamente en páginas diferentes que habrían sido impresas con el mismo tipo; letras parecidas, pero no tanto, impresas con tipos distintos pero provenientes de la misma matriz; y letras poco parecidas, impresas con tipos provenientes de matrices diferentes. No hemos abordado aquí esta posibilidad, más compleja, pero creemos que sería factible realizarla en el futuro. La propuesta de esta investigación se puede ampliar a otros libros impresos de la época, como por ejemplo el *Sibyllenbuch*, los Donatos, el Calendario Turco o el *Catholicon* entre otros. Véase no obstante el estudio con otras páginas del Evangelio de Mateo, iniciado en el apartado final 7.9.

El estudio tendrá un sesgo conservador. Es decir, se pedirá a los datos que nos den una evidencia suficiente de *la existencia de más de una matriz original*, ya que llegar a tal conclusión nos lleva a un replanteo de las tesis actualmente más aceptadas.

7. Clasificación de tipos en la Biblia de 42 líneas

7.2 El *cluster analysis*

El *cluster analysis* comprende una cierta variedad de métodos que pretenden obtener agrupaciones razonables de un conjunto de objetos, en base a las semejanzas y diferencias entre ellos. Cada grupo se denomina habitualmente *cluster*, y una agrupación particular, en que cada objeto ha sido asignado a un determinado cluster, se denomina un *clustering*. En español se usan a veces las expresiones "conglomerados" y "partición" en vez de las voces inglesas "cluster" y "clustering"; no obstante, mantendremos aquí casi siempre estas últimas, que son también de uso común en la literatura técnica en español.

Es de destacar que no es necesario que los objetos posean una característica cuantitativa relevante para realizar un cluster analysis. Es suficiente disponer de datos cuantitativos que se refieran a parejas de objetos y que representen la "distancia" entre uno y otro objeto. Por ejemplo, la superficie cubierta por la tinta de una letra impresa es una característica cuantitativa intrínseca de esa letra impresa, y se podría medir. Si quisiéramos agrupar letras según la superficie que cubren con la tinta, podríamos hacerlo en base a los valores de esas superficies. En nuestro caso, esta característica es irrelevante, y en cambio sí podremos establecer una distancia entre letras adecuada para nuestro propósito.

Idealmente, queremos que los clusters que formemos sean *cohesionados* (es decir, que las diferencias entre los objetos de un mismo cluster sean pequeñas) y que estén *aislados* (o sea que las diferencias entre objetos de distintos clusters sean grandes). Esa es exactamente la idea y la utilidad del cluster analysis. Pero hay diversas maneras, todas ellas razonables, de implementar esta idea. La elección de una u otra manera debe hacerse según cada caso y depende del tipo de conclusión que se aspire a obtener o la pregunta que se pretenda contestar. En el Apartado 7.4.2 se expondrán las diversas posibilidades y se justificará nuestra elección concreta.

El cluster analysis, si bien es nuestra principal herramienta, se complementará en dos direcciones, para el caso de las letras a, en el cual se aplica de un modo aproximado. Por un lado, se usarán los datos aportados por la medición de un pequeño conjunto de letras impresas a partir tipos nuevos, provenientes de una misma matriz. Este conjunto nos servirá como grupo de control para saber qué variabilidad puede esperarse de letras impresas con origen matricial común (los "errores aleatorios" arriba citados) (Apartado 7.6.1). Por otro lado, se usarán tests estadísticos de detección de *outliers* (datos extremos) para confirmar, con una probabilidad de equivocación controlada y pequeña, que dos

clusters distintos obtenidos mediante el cluster analysis deben considerarse efectivamente distintos. (Apartado 7.6.2).

Una exposición clásica del cluster analysis puede encontrarse por ejemplo en Gordon, [2]. Un planteo más moderno, como parte de la teoría llamada "unsupervised learning" (aprendizaje no supervisado), puede verse en el capítulo 14 de Friedman–Hastie–Tibshirani, [1]. La referencia básica sobre outliers es Barnett–Lewis [3].

Referencias

[1] Jerome Friedman, Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning* , Springer, 2001.
[2] A. D. Gordon. *Classification*, Chapman and Hall / CRC, 1999.
[3] V. Barnett, T. Lewis. *Outliers in statistical data*, Wiley, 1984.

7. Clasificación de tipos en la Biblia de 42 líneas

7.3 Cómo se mide la distancia entre letras

Dadas dos letras, se quiere tener una cuantificación de la diferencia entre ellas, es decir, una medida de la *distancia* o *disimilaridad* entre letras. Los términos "disimilaridad" y "distancia" pueden considerarse sinónimos, y los usaremos indistintamente[1].

El aparato de medida usado es el QVA-200 de Mitutoyo [6], con una resolución de captación de 0.0001 mm, y equipado con los programas QVPack y FormPack. Las letras se han medido a partir de un microfilm de seguridad del Nuevo Testamento de la Bíblia de 42 líneas[2] que se encuentra en la biblioteca de la Universidad de Sevilla.

Para estimar el error de medida atribuible al aparato y al proceso de medición en sí, se midió la distancia de un escaneo consigo mismo y las distancias fueron del orden de 10^{-27} (Fig. 129). Dos escaneos diferentes de la misma letra dieron lugar a una distancia de 10^{-16}. Como se verá, estos valores están muchos órdenes de magnitud por debajo de los valores relevantes al medir la disimilaridad entre dos letras impresas distintas, por lo que el proceso de medición hay que considerarlo estable y fiable en este aspecto. Este proceso es el que se describe a continuación:

El aparato de medida escanea cada letra y determina su *contorno*. Un contorno esta formado por puntos y por segmentos que unen los puntos, formando un curva cerrada, rectilínea a trozos. Algunas letras poseen espacios vacíos interiores, por lo que el contorno puede estar formado por varias componentes conexas.

Para comparar los contornos de dos letras con el programa QVPak, se asigna a uno de ellos el papel de "medido" y al otro el de "nominal", según la terminología del propio programa. La disimilaridad entre un medido y un nominal se determina mediante los pasos siguientes:

1. Los dos contornos ocupan inicialmente una posición cualquiera en el plano.

2. El contorno medido se resitúa de modo que coincidan los centros de gravedad[3] de los puntos que forman ambos contornos. Éstos quedan así aproximadamente superpuestos.

3. En esta posición, se calcula un valor provisional de la disimilaridad entre las letras de la manera siguiente:

a) Se mide la *distancia euclídea* (es decir, la distancia ordinaria), elevada al cuadrado, desde cada punto del perfil nominal al perfil medido (concretamente, al punto más cercano del segmento más cercano).

b) Se suman todas estas distancias al cuadrado para todos los puntos del nominal.

4. En función de ese valor y de la posición actuales se resitúa el contorno medido en otra posición en la que se espera obtener un valor más pequeño.

5. Se repiten los pasos 3 y 4 las veces que sean necesarias hasta que la mejora (disminución) de la suma de cuadrados entre pasos sucesivos es despreciable.

6. En la posición final resultante admitimos que tenemos las letras lo mejor superpuestas posible, y por tanto el valor actual de la suma de cuadrados es el mínimo posible. Lo llamamos *residual*.

7. Como valor de disimilaridad entre letras no conviene tomar directamente este residual, porque influiría en el resultado el número de puntos del perfil nominal, que puede ser distinto en cada comparación. Tomaremos el *residual medio*, obtenido dividiendo el residual que nos da el paso 6 por el número de puntos del nominal.

8. Se intercambian los papeles del contorno medido y el contorno nominal, se repiten los pasos del 1 al 7 y se toma como valor de disimilaridad definitivo la media de los residuales medios obtenidos.

Como he ha dicho, tomando el residual medio en el paso 7 se evita que el resultado dependa del número de puntos del contorno nominal. Esto nos permitirá más tarde comparar entre sí de manera homogénea todas las letras, independientemente del número de puntos exacto de que conste el nominal en cada caso.

El motivo de intercambiar papeles en el paso 8 es que el resultado varía ligeramente dependiendo de qué perfil se toma como nominal y cuál se toma como medido, como es esperable del hecho de que en el paso 3 los dos contornos no juegan el mismo rol.

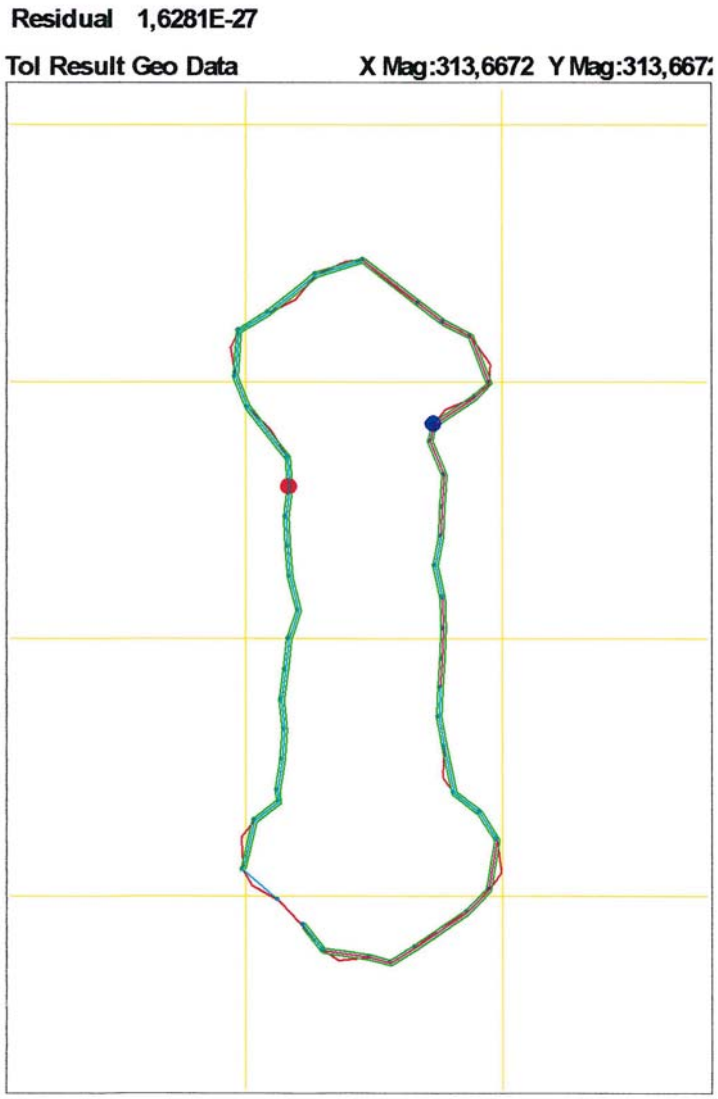


Fig. 129 Comparación de i4 consigo misma.

Referencias

[6] Mitutoyo, http://www.mitutoyo.com.

Notas

1. Técnicamente, el término "distancia", en el sentido matemático habitual, debe cumplir la desigualdad triangular: la distancia entre dos objetos ha de ser menor que la suma de las distancias de esos objetos a un tercero. Esto no lo cumple la medida de disimilaridad que vamos a usar.
2. Gutenberg, Johannes. *Biblia Latina*. Moguntiae, (c. 1454- agosto, 1456). Alemania. Depositada en el Fondo Antiguo de la Biblioteca de la Universidad de Sevilla.
3. El centro de gravedad de un conjunto de puntos con coordenadas (x_1, y_1), ..., (x_n, y_n) es el punto

$$\frac{1}{n}(x_1 + \ldots + x_n\ ), \frac{1}{n}(y_1 + \ldots + y_n\ )$$

7. Clasificación de tipos en la Biblia de 42 líneas

7.4 *Cluster analysis* para la letra i

7.4.1 Tabla de disimilaridades

Las 10 letras i impresas que aparecen en las líneas impares de la primera página del Evangelio de San Mateo se han comparado con el procedimiento descrito en el apartado 7.3, todas con todas, obteniéndose una tabla simétrica de disimilaridades, que es el dato inicial adecuado para proceder al cluster analysis. Ya que no son verdaderas distancias[4], es imposible representar visualmente las letras como puntos en un espacio coordenado, de la dimensión que sea. No obstante, mediante la técnica del *Multidimensional Scaling* (véase el Apartado 7.7), se obtendrá una cierta visualización de la posición relativa aproximada entre las letras, en espacios de dos y de tres dimensiones.

Las dimensiones originales de las letras en el microfilm son del orden de 0.285 mm (para la altura *x*), y dan lugar a disimilaridades entre ellas que son del orden de 10^{-5} mm^2. Para visualizar y tratar más cómodamente los valores obtenidos, éstos han sido multiplicados por 10,000, resultando las disimilaridades de la Tabla 44.

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
i1		0.2329	0.3518	0.2310	0.1912	0.3213	0.2179	0.2652	0.2590	0.3929
i2	0.2329		0.2097	0.0283	0.0701	0.0801	0.1139	0.0425	0.0835	0.0907
i3	0.3518	0.2097		0.1750	0.1638	0.3713	0.0670	0.1901	0.1232	0.3290
i4	0.2310	0.0283	0.1750		0.0895	0.0756	0.0902	0.0512	0.1026	0.0926
i5	0.1912	0.0701	0.1638	0.0895		0.1541	0.1219	0.1285	0.1028	0.2014
i6	0.3213	0.0801	0.3713	0.0756	0.1541		0.1467	0.0560	0.1324	0.0645
i7	0.2179	0.1139	0.0670	0.0902	0.1219	0.1467		0.0977	0.0900	0.2186
i8	0.2652	0.0425	0.1901	0.0512	0.1285	0.0560	0.0977		0.0718	0.0719
i9	0.2590	0.0835	0.1232	0.1026	0.1028	0.1324	0.0900	0.0718		0.1717
i10	0.3929	0.0907	0.3290	0.0926	0.2014	0.0645	0.2186	0.0719	0.1717	

Tabla 44. Tabla de disimilaridades para las i.

En los Apartados siguientes estudiaremos la clasificación de las letras i a partir de esta tabla. Al mismo tiempo, introduciremos la teoría del "cluster analysis" necesaria que permita justificar nuestros métodos y resultados.

Los programas que se han elaborado para el análisis que sigue se han escrito en el lenguaje *R*. Se trata de un lenguaje especializado en el tratamiento de datos que actualmente se está convirtiendo en un estándar en las aplicaciones científicas. Se distribuye como "open source", junto con un entorno gráfico, a través de *The R Project for Statistical Computing* [4]. Existe también una implementación comercial, llamada *S-plus* [5].

7.4.2 Cómo valorar cada posible *clustering*

Como se mencionó en el Apartado 7.2, queremos que los clusters sean *cohesionados* y *aislados*. Supongamos que tenemos una cantidad N de objetos (en nuestro caso, tenemos $N = 10$ letras i). Y supongamos que queremos subdividir los N objetos en un cierto número K de clusters. (En realidad, no queremos fijar de entrada el número de clusters, pero supongamos de momento que lo hacemos. Más tarde volveremos a la discusión sobre el número de clusters.)

Ahora queremos encontrar el clustering formado por K clusters que sea el mejor en términos de las propiedades de cohesión y aislamiento mencionadas. Para poder hablar de "el mejor" clustering, o para poder comparar dos clustering candidatos, hay que establecer un valor numérico (un "coste" o "penalización") a cada clustering y escoger entre todos los posibles clusterings aquel que tenga el mejor valor (el menor "coste").

No es evidente como transformar las cualidades de cohesión y aislamiento en valores numéricos. De hecho, hay diversas maneras razonables de hacerlo. En cada situación es necesario escoger entre ellas la que parezca más adecuada al problema concreto que se quiere resolver.

Veamos algunas valoraciones posibles para motivar la elección que haremos. En primer lugar, hay que valorar el coste de cada cluster concreto, y luego combinar los costes de los diferentes clusters de un clustering dado. Ambas cosas se pueden hacer de varias maneras. Para la primera (cuál es el coste de un cluster concreto), típicamente se utiliza alguna de las siguientes cantidades:

1. El máximo de las disimilaridades entre objetos del cluster.

2. La suma de las disimilaridades entre objetos del cluster.

3. Para cada objeto, se mide la suma de todas las disimilaridades entre él y los demás objetos del cluster. De entre todas estas cantidades, se toma la más pequeña.

También se utilizan a veces valoraciones basadas en el mínimo o la suma de disimilaridades entre los objetos del cluster y los que no son del cluster. Añadiremos una variante del criterio 3, que será la que finalmente usaremos; véase la justificación más adelante.

4. Para cada objeto, se mide la media de todas las disimilaridades entre él y los demás objetos del cluster. De entre todas estas cantidades, se toma la más pequeña.

Una vez seleccionada una manera de valorar cada cluster, hemos de combinar las valoraciones de los distintos clusters. Esto se puede hacer de dos modos razonables:

a. Sumando los valores de todos los clusters.
b. Tomando el máximo de los valores de todos los clusters.

Los métodos 3 y 4 tienen la ventaja de que distinguen un objeto particular de cada cluster, que es aquel para el cual la suma o la media de las disimilaridades con los demás objetos es la más pequeña. Esto permite pensar en los demás objetos como situados "alrededor de él". En nuestro caso, permite tomar una letra como "modelo" y pensar que las demás son "variantes" del modelo, representando, en teoría, letras impresas que posiblemente provienen de la misma matriz que el modelo pero de distintos tipos y/o con errores de impresión.

Entre las posibilidades (a) y (b), nos inclinamos por la segunda, ya que con la primera puede haber clusters con una valoración muy alta junto con otros con una valoración muy baja. Tomar el máximo como valor a minimizar tiende a uniformizar los clusters en este sentido, lo cual nos parece más razonable y conservador en la situación que estamos estudiando.

Una vez decidida la utilización del criterio (b) para la combinación de valoraciones de todos los clusters, la nueva alternativa propuesta 4 es mejor que la idea estándar 3, como muestra el siguiente ejemplo: Supongamos que para las letras i queremos comparar el clustering

[{1}, {2, 3, 4, 5, 6, 7, 8, 9, 10}]

en el que la letra 1 está aislada y los demás estén reunidas en el mismo clúster, con el clustering

[{1, 10}, {2, 3, 4, 5, 6, 7, 8, 9}].

Con la combinación 3-b, el primero tiene un coste de 0.7050 y el segundo de 0.6124, siendo por tanto mejor esta última opción. Pero esto no parece razonable, porque la letras 1 y 10 son las que están más alejadas entre sí. Lo que ha sucedido aquí es que en el primer clustering tenemos un cluster con valor 0 (por tener sólo un elemento, la cohesión interna es total), y otro con valor 0.7050; en el segundo clustering, el primer cluster pasa a valer 0.3929, que es la distancia entre las letras 1 y 10, y el segundo disminuye a 0.6124. El máximo, por tanto, ha disminuido de 0.7050 a 0.6124, y encontramos que el segundo clustering es mejor.

Este inconveniente tendrá tendencia a aparecer con la combinación de criterios 3-b cuando tengamos clusters con una cantidad muy desigual de elementos. Con la alternativa 4-b, el hecho de que un cluster tenga muchos elementos y otro tenga pocos no influye, porque se toma la media. En el mismo ejemplo anterior, ahora el primer clustering tiene un coste de 0.0881, mucho mejor que el segundo, que lo tiene de 0.3929.

En resumen, creemos que la combinación de criterios 4-b es la más adecuada para este estudio, dando lugar a que:

· Los clusters podrán representarse gráficamente en forma de estrella, con una letra modelo en su centro, y las demás letras del cluster a su alrededor, como variantes del modelo.
· No habrá clusters muy cohesionados conviviendo con clusters poco cohesionados.
· No habrá, en la medida de lo posible, letras muy alejadas del centro de la estrella, conviviendo con otras muy cercanas.

7.4.3 Cómo encontrar el *clustering* óptimo

Una vez decidido el criterio con el que se va a dar un valor a cada posible clustering, tenemos la cuestión de cómo encontrar la mejor partición según ese criterio.

Se puede intentar construir todas las posibles particiones, valorarlas, y elegir la mejor (u obtener una ordenación de mejor a peor de todas ellas). El problema es que el número de particiones posibles puede ser demasiado grande. Con un programa y un ordenador muy rápidos, capaces de generar y evaluar 100 clusterings por segundo, el tiempo necesario para el cálculo con 10 y 21 objetos y varias cantidades de clusters sería el de la Tabla 45, al que habría que añadir el tiempo necesario para ordenar la lista resultante.

Número de objetos	Número de clusters	Clusterings posibles	Tiempo de cálculo
10	2	511	5 segundos
10	3	9 330	93 segundos
10	4	34 105	6 minutos
10	5	45 525	7 minutos
21	2	1 048 575	3 horas
21	3	1 742 343 625	202 días
21	4	181 509 070 050	57 años
21	5	3 791 262 568 401	1202 años
21	6	26 585 679 462 804	8430 años

Tabla 45. Tiempo estimado de cálculo pera la valoración de todos los posibles clusterings.

Hemos usado este método directo y exacto para clasificar de la mejor manera posible (según el criterio 4-b explicado en el Apartado 7.4.2) las 10 letras i en dos y tres clusters. La mejor partición en dos clusters ha sido

{1} {2, 3, 4, 5, 6, 7, 8, 9, 10},

mientras que en tres clusters ha sido

{1} {3, 7} {2, 4, 5, 6, 8, 9, 10}.

Véase en el Apéndice 7.10 la lista de los veinte mejores clusterings en cada caso.

Es posible representar gráficamente estos clusterings mediante “estrellas”. El mejor clustering con dos clusters, que consiste en la letra 1 aislada y todas las demás alrededor de la letra 4 como modelo, puede representarse como se ve en la Fig. 130, en el caso de tres clusters se representa de manera análoga en la Fig. 131.

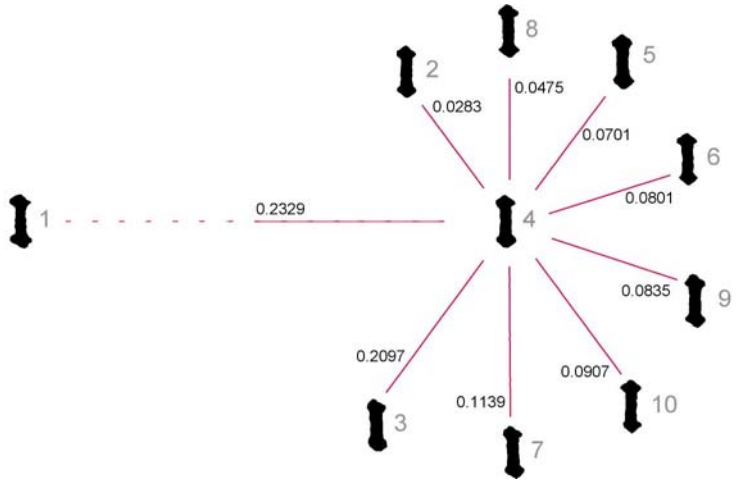


Fig. 130 El mejor clustering con dos clusters representado mediante estrellas.

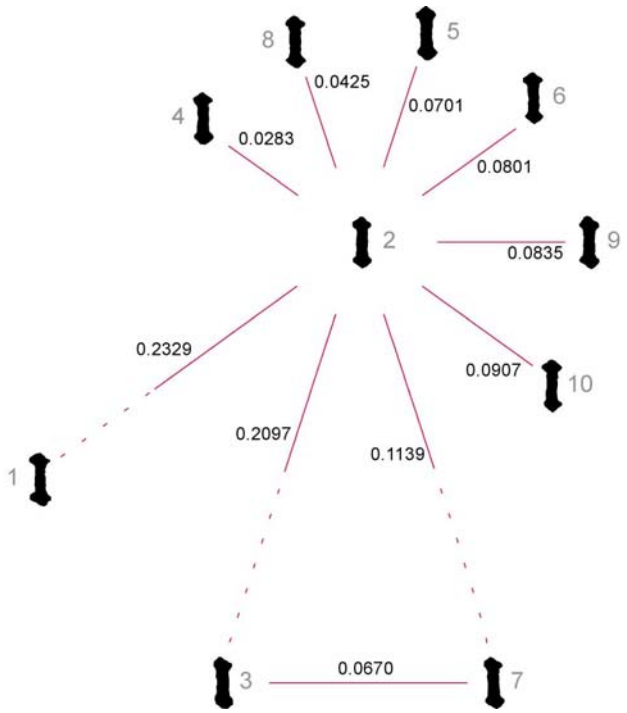


Fig. 131 El mejor clustering con tres clusters representado mediante estrellas.

En ambos gráficos, las letras pertenecientes al mismo cluster se han representado unidas mediante una línea continua a la letra modelo del cluster, con la disimilaridad anotada sobre ella. Los modelos de clusters distintos se han unido mediante una línea discontinua con el valor de la disimilaridad entre modelos también sobre ella. No obstante, las longitudes de las líneas no son proporcionales a la disimilaridad. Obsérvese también que, en el caso de clusters de dos letras exactamente, no hay ninguna razón para elegir una u otra como modelo.

Obsérvese que tanto en el caso de dos como en el de tres clusters, las distancias entre las letras modelo son mayores que todas las demás distancias. Esto no es una consecuencia necesaria del criterio

escogido para encontrar el clustering óptimo, sino que es una propiedad adicional que abunda en la bondad de estos clustering particulares.

Puede observarse un cambio en la letra modelo del cluster mayor, cuando pasamos de dos a tres clusters. Este cambio se ha producido al no tener en cuenta en el cálculo del segundo caso las letras i3 e i7, y no es un fenómeno extraño, ya que de hecho las letras i2 y i4 son muy parecidas (disimilaridad 0.0283, muy pequeña en comparación a todas las demás).

Podría también considerarse la mejor subdivisión en cuatro o más clusters, pero con tan pocos objetos creemos que no tiene sentido, y sería en todo caso arriesgada. Obsérvese que en el caso extremo de considerar cada letra aislada en su propio cluster, la propiedad de mayor separación entre modelos que en el interior de cada cluster (con respecto a su modelo) se cumpliría trivialmente. Pero para un tal clustering no puede haber ningún criterio de validación con sentido.

Para las 21 letras a, el método directo y exacto desarrollado aquí no es computacionalmente viable, y hemos recurrido a un método aproximado, que veremos en los apartados siguientes.

7.4.4 Número de clusters

La cuestión de cuántos clusters hay, o de si realmente hay más de uno, no tiene una solución universal. Nuestra filosofía se ha basado en las consideraciones siguientes:

· En primer lugar, queremos ser conservadores; es decir, preferimos equivocarnos afirmando que dos letras provienen de la misma matriz aunque en realidad provengan de matrices distintas, que equivocarnos diciendo que provienen de distinta matriz cuando en realidad provienen de la misma.

· En segundo lugar, nuestro objetivo básico no es determinar exactamente el número de matrices distintas usadas. Se trata de ver si hay suficiente evidencia de que hay más de una, más de dos, etc.

· En tercer y último lugar, no hay más remedio que tomar algunas decisiones *a priori* que pueden influir en el resultado final; son las decisiones de adoptar uno u otro criterio sobre qué es una partición óptima, o elegir el método aproximado que se va a emplear. Pero, *a posteriori*, una vez se tienen los resultados de los procedimientos de clustering, se reexaminarán éstos con un criterio adicional para confirmar el aislamiento de los clusters.

Referencias

[4] The R Project for Statistical Computing, http:\\www.r-project.org.
[5] S-Plus, http://www.insightful.com.

Notas

4. Véase la nota 1.

7. Clasificación de tipos en la Biblia de 42 líneas

7.5 *Cluster analysis* aproximado

No hay ninguna manera de calcular la mejor partición que sea esencialmente más eficiente que la enumeración exhaustiva que hemos aplicado a las i. Por lo tanto, si por razones de la dimensión del problema no se puede hacer esta enumeración, hay que conformarse con algún método aproximado, que nos proporcione una solución razonablemente buena.

El método aproximado que usaremos responde a un procedimiento heurístico llamado *aglomerative hierarchical clustering* (clustering jerárquico aglomerativo). Con este método se hace en realidad una jerarquía de clusterings, de modo que a posteriori hay que decidir con qué clustering de la jerarquía nos quedamos.

Se empieza considerando que cada objeto constituye un cluster por sí solo, y se van construyendo clusters cada vez mayores. En cada paso, se unen dos clusters en un cluster mayor. Concretamente se unen los dos clusters que se encuentren más cercanos.

Esto nos lleva a tener que definir una noción de disimilaridad entre clusters, y aquí también hay una cierta arbitrariedad en la elección. Hay tres posibilidades razonables:

1. La disimilaridad entre dos clusters es la disimilaridad entre sus objetos más semejantes. (Esta opción se conoce en la literatura como *single linkage*.)

2. La disimilaridad entre dos clusters es la disimilaridad entre sus objetos más distintos (conocida como *complete linkage*).

3. La disimilaridad entre dos clusters es la media de disimilaridades entre los objetos de los dos clusters (denominada *average linkage*).

Antes de atacar la clasificación de las letras a, ilustraremos el clustering jerárquico aglomerativo resultante de las tres posibilidades anteriores con las letras i ya estudiadas con un método exacto, y discutiremos las bondades e inconvenientes de cada definición de disimilaridad entre clusters.

7.5.1 El “single linkage”

Con cualquiera de las tres posibilidades se puede hacer una representación gráfica del proceso aglomerativo. Esta representación se llama *dendrograma*. Por ejemplo, el procedimiento de “single linkage” para las 10 letras i produce el dendrograma de la Fig. 132.

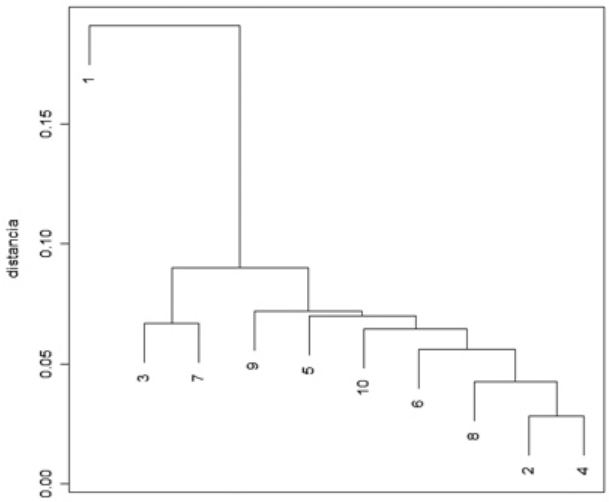


Fig. 132 Dendrograma para la letra i con el método “single linkage”.

La tabla siguiente indica el orden (números entre corchetes) en que se forman los clusters, y cuál es la distancia entre los que se juntan. Las dos columnas centrales indican qué clusters se juntan (un guión delante de un número indica que es el objeto con ese número; si no hay guión, se refiere al cluster que se formó en la línea de ese número.)

```
[1]   -2   -4  0.0283
[2]   -8    1  0.0425
[3]   -6    2  0.0560
[4]  -10    3  0.0645
[5]   -3   -7  0.0670
[6]   -5    4  0.0701
[7]   -9    6  0.0718
[8]    5    7  0.0900
[9]   -1    8  0.1912
```

Siguiendo la tabla y el dendrograma, vemos que en primer lugar hemos aglomerado los dos objetos que más se parecían, i2 e i4, cuya distancia es de 0.0283; seguidamente, la disimilaridad más pequeña se produce entre el i8 e i2, y es de 0.0425, por lo que añadimos i8 al cluster que habíamos formado en la primera línea de la tabla; seguidamente son i6 e i8 las letras que se parecen más, lo que nos lleva a

juntar i6 con el cluster formado en la línea 2; después i10 se añade también al mismo cluster. En el siguiente paso la disimilaridad más pequeña es la que hay entre las letras i3 e i7. Después i5 e i9 se añaden al primer cluster. Luego ambos clusters se funden en uno, y finalmente i1 se añade al grupo.

Ahora, a la vista del dendrograma, se debería decidir cuántos y cuáles clusters nos indica que hay. Cuando estudiemos las letras a con este método aproximado, usaremos información adicional que nos ayudará en este propósito. Pero en este momento queremos sólo observar hasta qué punto los métodos aproximados dan resultados semejantes al método exacto que hemos empleado para las i.

Concretamente, vemos en el dendrograma que claramente la letra 1 debe estar aparte de las demás, y se observa también una separación clara entre el grupo {3, 7} y el resto de las letras, aunque la separación no es tan evidente como con la letra 1. Esto refleja bastante bien de manera gráfica los resultados que hemos obtenido al imponer dos y tres clusters respectivamente. Asimismo, buscar una separación en más de tres clusters no se vería apoyada por el dendrograma.

El método single linkage es muy conservador, en el sentido de que puede fácilmente considerar en un mismo cluster objetos muy alejados, pero que están unidos por una larga cadena de objetos, cada uno similar al siguiente. Se le conoce informalmente como el método "los amigos de mis amigos son mis amigos".

7.5.2 El método "complete linkage"

El método "complete linkage" opta por la heurística contraria al single linkage y esto lo hace poco conservador: Tiende a no juntar en el mismo cluster objetos que podrían, tal vez, ir separados. Los clusters que produce tienden a ser cohesionados pero poco aislados. Con el "complete linkage" obtenemos el dendrograma de la Fig. 133.

La secuencia de formación de los clusters es:

[1]	-2	-4	0.0283
[2]	-8	1	0.0512
[3]	-6	-10	0.0645
[4]	-3	-7	0.0670
[5]	2	3	0.0926
[6]	-5	-9	0.1028
[7]	4	6	0.1638
[8]	-1	7	0.3518
[9]	5	8	0.3929

Después de juntar i2 e i4, como en el caso del "single linkage", también aquí el siguiente paso es juntar el 8 con ellos, pero ahora no porque la distancia de i8 a i2 (0.0425) sea la siguiente más pequeña, sino porque la distancia de i8 a i4 (0.0512) lo es. Después i6 forma grupo con i10, porque la distancia entre ellos es de 0.0645, mientras que la distancia de i6 al grupo {i2, i4, i8} es de 0.0801 (en el "single linkage" era de 0.0560). El procedimiento sigue avanzando análogamente hasta obtener un sólo cluster.

El dendrograma sugiere unos resultados que concuerdan menos que el "single linkage" con los que hemos obtenido con el método exacto.

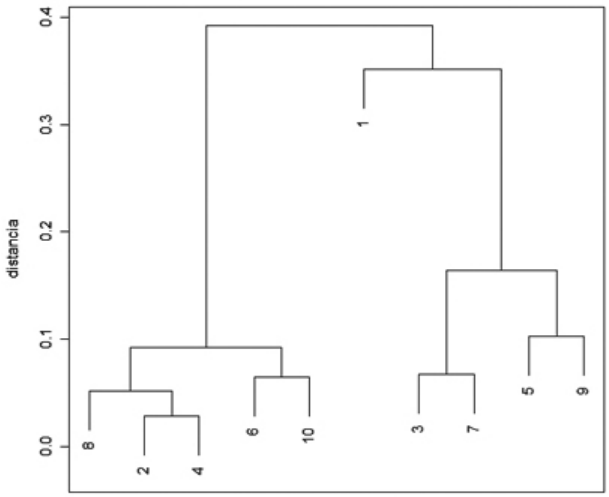


Fig. 133 Dendrograma para la letra `i` con el método "complete linkage".

7.5.3 El método "average linkage"

El método "average linkage" está a medio camino entre los anteriores. Con este método obtenemos el dendrograma de la Fig. 134.

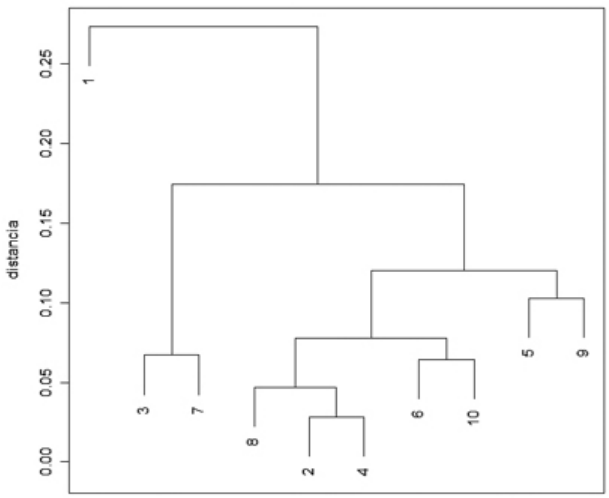


Fig. 134 Dendrograma para la letra `i` con el método "average linkage".

La secuencia de formación de clusters es:

```
[1]    -2    -4    0.0283
[2]    -8     1    0.0469
[3]    -6   -10    0.0645
[4]    -3    -7    0.0670
[5]     2     3    0.0778
[6]    -5    -9    0.1028
[7]     5     6    0.1205
[8]     4     7    0.1744
[9]    -1     8    0.2737
```

Observamos que en nuestro caso se parece más a lo que hemos obtenido con el "single linkage", en el sentido de que los cortes más naturales producen las mismas particiones en dos y tres clusters que en el "single linkage". Asimismo, la separación del grupo {3, 7} parece aquí más evidente, por lo que este método aproximado nos parece el que da resultados más parecidos al método exacto correspondiente al criterio 4-b que hemos aplicado. Tendremos en cuenta esta conclusión al estudiar las veintiuna letras a.

7. Clasificación de tipos en la Biblia de 42 líneas

7.6 *Cluster analysis* para la letra a

Se ha procedido de manera análoga al Apartado 7.5 para la comparación de las veintiuna a de la misma página. La tabla correspondiente de disimilaridades se encuentra en el Apéndice 7.10.2.

7.6.1 Dendrogramas para la letra a

Por las razones apuntadas en los Apartados 7.1 y 7.5, haremos una clasificación inicial de las 21 a usando un método aproximado de clustering. Concretamente, usaremos el clustering jerárquico aglomerativo con la distancia "average linkage", que es el método que nos ha dado resultados más parecidos al clustering exacto en el caso de las i.

El dendrograma resultante se observa en la Fig. 135.

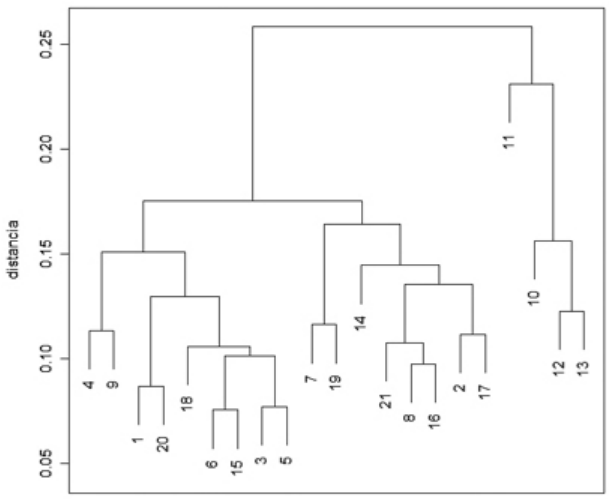


Fig. 135 Dendrograma para la letra a con el método "average linkage".

El orden en que están ordenadas las sucesivas agrupaciones en sentido horizontal no tiene ningún significado especial. Aquí se ha seguido el criterio de colocar a la izquierda el cluster más compacto, es decir, aquel de los dos que se ha formado a un nivel más bajo.
La correspondiente lista de formación de clusters es

```
[1]      -6     -15     0.0754
[2]      -3      -5     0.0769
[3]      -1     -20     0.0868
[4]      -8     -16     0.0972
[5]       1       2     0.1012
[6]     -18       5     0.1056
[7]     -21       4     0.1074
[8]      -2     -17     0.1114
[9]      -4      -9     0.1133
[10]     -7     -19     0.1161
[11]    -12     -13     0.1226
[12]     3       6     0.1296
[13]     7       8     0.1351
[14]    -14     13     0.1445
[15]     9      12     0.1506
[16]    -10     11     0.1562
[17]     10     14     0.1640
[18]     15     17     0.1752
[19]    -11     16     0.2309
[20]     18     19     0.2584
```

El dendrograma nos da una idea gráfica sobre qué letras pueden ser más parecidas a otras, pero queda pendiente determinar a partir de él cuántos y cuáles clusters creemos que hay.

Si no tuviéramos ninguna otra información al alcance, sólo podríamos conjeturar el número y composición de los clusters basándonos en los saltos de distancias entre aglomeraciones consecutivas, y "cortar el dendrograma" por allí donde hubiera saltos grandes de magnitud. Por ejemplo, el salto de 0.1752 a 0.2309 entre las líneas [18] y [19] nos daría una división en tres clusters que parece clara.

Pero sí disponemos de una información adicional para determinar el número y composición de los clusters. Esta información la proporciona un conjunto de mediciones de disimilaridades que se ha realizado a partir del escáner de la impresión de unos *tipos nuevos* provenientes de una *póliza francesa*[5] que pertenece al Fondo Tipográfico de la Fundición Bauer-Neufville[6], por lo que podemos estar seguros que nunca se han usado y que provienen todas de la misma matriz; éstas harán el papel de "grupo de control". Si pretendiéramos dividirlas en clusters todas ellas deberían por tanto formar parte del mismo cluster. Comparando entre sí estas letras de la misma forma que se ha hecho con las letras microfilmadas del Evangelio de Mateo, veremos qué disimilaridades podemos esperar atribuibles sólo a los errores de construcción de los tipos y a los errores de impresión.

El tamaño real de los tipos nuevos es distinto al de las letras microfilmadas, pues corresponden a la tipografía Venus 16 pts, ancha súper negra (ref. 2667, Hauter Didot) por lo que los valores absolutos de las disimilaridades no pueden usarse directamente para comparar las unas con las otras. Tenemos que obtener una base común de comparación que no se vea influida por esta diferencia de tamaño. Nos parece natural proceder del siguiente modo:

1. Agrupar los tipos nuevos alrededor de una letra modelo, con el mismo procedimiento que hemos aplicado a las letras `i` en el Apartado 7.4 (todas ellas en el mismo cluster, naturalmente), obteniendo un conjunto de distancias a la letra modelo.
2. Calcular el cociente entre la distancia mayor y la distancia menor al modelo, como medida de dispersión de estas distancias, que es un valor totalmente independiente de las unidades en que se calcule.

Se ha aplicado este procedimiento a varias letras, concretamente las letras `m`, `i`, `a`, `o`, con cinco especimenes de cada una. Véase el resultado en la Tabla 46.

	Distancias al modelo				media	max/min
m-nueva	5.9731,	3.9409,	5.5055,	3.9967	4.8540	1.5156
i-nueva	3.2364,	5.9557,	2.8876,	3.9188	3.9996	2.0625
a-nueva	7.4534,	4.8179,	4.6993,	3.7450	5.1789	1.9902
o-nueva	1.7965,	1.8714,	2.0286,	1.7940	1.8726	1.1307

Tabla 46. Distancias obtenidas con los tipos nuevos.

La división entre la máxima distancia y la mínima distancia al modelo de los tipos nuevos nos da una medida de la variabilidad que podemos esperar en este tipo de letras. Vemos que los valores de esta división oscilan entre 1.1307 y 2.0625.

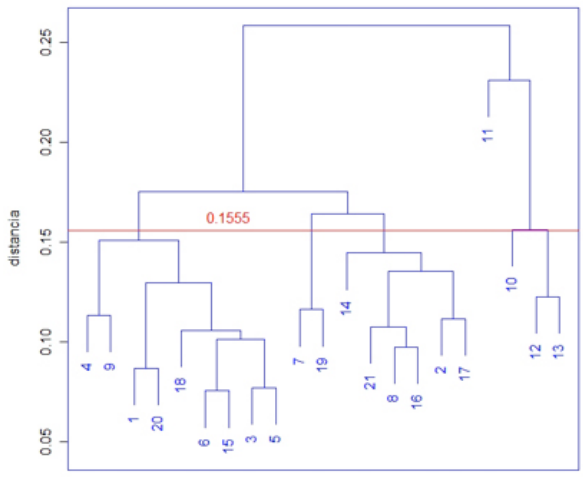


Fig. 136 Distancia por debajo de la cuál no subdividiremos clusters.

Podemos usar esto del siguiente modo. Para las a de la primera página del Evangelio de Mateo, tomemos las dos letras más cercanas y declarémoslas, evidentemente, como pertenecientes al mismo cluster. Esa distancia mínima es 0.0754, y se produce entre a6 e a15. Según los datos obtenidos con los tipos nuevos, las letras a distancias de hasta 2.0625 veces esta distancia mínima (es decir, distancias de hasta 0.1555) pueden perfectamente pertenecer al mismo cluster.

Por lo tanto, podemos trazar una línea horizontal en el dendrograma a la altura 0.1555, y declarar todos los clusters que quedan agrupados debajo de la línea como clusters que no vamos a subdividir (véase la Fig. 136).

Así, hemos obtenido seis clusters (Tabla 47). Podemos ahora construir las estrellas correspondientes a estos clusters, obteniendo unas letras modelo, y las correspondientes distancias de las demás letras a la letra modelo:

número	cluster	modelo	media de las distancias	mínima, máxima	máx/mín
1	{4, 9, 1, 20, 18, 6, 15, 3, 5}	5	0.1114	0.0769, 0.1468	1.9090
2	{7, 19}	7, 19	0.1161		
3	{14, 21, 8, 16, 2, 17}	8	0.1137	0.0972, 0.1347	1.3858
4	{11}	11	–		
5	{10}	10	–		
6		12, 13	0.1226		

Tabla 47. Clustering preliminar para las letras a.

La conclusión hasta este momento es que *el número máximo de clusters que obtenemos de las veintiuna letras a es 6*, puesto que subdividir cualquiera de ellos sería análogo a declarar que los tipos nuevos estudiados provenían de más de una matriz.

7.6.2 Validación de los clusters obtenidos

Ahora nos interesa ver si en realidad hemos ido demasiado lejos y algunos de los clusters anteriores deberían aglomerarse en clusters mayores, siguiendo el criterio conservador de no declarar dos letras como provenientes de matrices distintas si no hay mucha evidencia que apoye esta conclusión. Téngase también en cuenta que el dendrograma es el resultado de un método aproximado, con una elección un tanto arbitraria del criterio con que se construye; por lo tanto, es de todo punto necesario validar los clusters que tenemos ahora establecidos con algún criterio alternativo que refleje el espíritu conservador.

Para esta validación, la herramienta a usar será un test estadístico para la detección de "outliers". En general, un *outlier* en un conjunto de datos numéricos es un dato muy diferente de los demás, que parece ser debido a un error tipográfico o de medida, o simplemente que *no debería estar ahí*, porque no pertenece realmente a ese conjunto de datos.

¿Cómo se detecta la presencia de un outlier? En principio, a la vista solamente de los valores numéricos, y a no ser que haya algún otro motivo para sospechar la presencia de un outlier, debe suponerse que todos lo datos son legítimos. Esa es la hipótesis conservadora, y que se va a mantener hasta el final, *siempre y cuando no haya razones poderosas para inclinarse por la hipótesis contraria,* razones aportadas por los propios datos numéricos.

En general, todos los tests estadísticos proceden de la manera descrita: Se parte de una hipótesis "conservadora" (es decir, la menos arriesgada o la que representa el *statu quo* o el conocimiento establecido hasta el momento), que sólo será desestimada si los datos numéricos aportan suficiente evidencia en su contra. Esa evidencia puede medirse mediante la probabilidad de equivocarse cuando se rechaza la hipótesis.

Habitualmente se fija de antemano un máximo tolerable a la probabilidad de equivocarse al rechazar la hipótesis formulada. Un valor muy común es el 5%; para estudios con alto riesgo al rechazar la hipótesis (por ejemplo en el ámbito farmacéutico) se usan valores del 1% o incluso menores; en estudios en que el investigador adopta una actitud más arriesgada se puede utilizar el 10%. Una vez fijado este máximo tolerable de riesgo de equivocación (digamos que del 5%), las conclusiones del test de hipótesis sólo pueden ser dos:

a. Rechazar la hipótesis formulada, admitiendo que la conclusión puede ser errónea, pero que la probabilidad de que lo sea es menor del 5%.
b. Aceptar la hipótesis. En tal caso, la probabilidad de estar equivocándose no está controlada.

Cómo se ve de las dos posibles conclusiones, la hipótesis inicial de trabajo parte con ventaja, puesto que sólo se la rechazará si hay suficiente evidencia numérica en su contra; y en cambio ella no tiene por qué aportar evidencias numéricas en su favor.

Aplicaremos estas ideas del siguiente modo: Tomemos por ejemplo el cluster número 3. El modelo es a8 y podríamos dibujar por tanto una estrella con centro en a8 y las correspondientes distancias a las demás letras. Estas distancias son, concretamente, ordenadas de menor a mayor:

0.0972 0.1041 0.1105 0.1221 0.1347

Tomemos ahora otra letra cualquiera, ajena al cluster 3, y anotemos su distancia al modelo a8. Por ejemplo, a1 se encuentra a distancia 0.1536 de a8. ¿Es esta distancia exageradamente grande para colocar a1 también en el cluster 3? O, por el contrario, no desentona con las demás, y ¿podría ser casual que el valor sea algo mayor?

La hipótesis conservadora y a la que por tanto hay que dar validez *a priori* es la de que la distancia entre a1 y a8 no es exagerada y en consecuencia a1 debería pertenecer también al cluster 3.

El test estadístico más habitual para poner a prueba esta hipótesis es el llamado *test de Dixon*. Si $x_1,\ldots,x_n$ son los datos que se tienen, ordenados de menor a mayor, *incluyendo* el dato sospechoso, el test de Dixon consiste en calcular

$$\frac{x_n - x_{n-1}}{x_n - x_1},$$

esto es, el último valor (el sospechoso) menos el penúltimo, dividido por la diferencia entre el mayor y el menor valor. Intuitivamente está claro que, si este cociente nos da un valor grande, significa que hay mucha diferencia entre el penúltimo y el último valor (relativamente a la dispersión de todos los datos) y esto nos induce a pensar que ese último valor es un outlier y por lo tanto nos lleva a rechazar la hipótesis de trabajo. Si, por el contrario, el valor es pequeño, entonces la diferencia entre penúltimo y último valor es pequeña y no hay razón para declarar a éste como outlier.

Por ejemplo, si queremos ver si la letra **a**1 pertenece al cluster 3, calcularemos el valor

$$\frac{0.1536-0.1347}{0.1536-0.0972}=0.3551$$

Cómo determinar si el valor 0.3351 obtenido es demasiado grande o no lo es? Afortunadamente, bajo ciertas condiciones que en nuestro caso se cumplen (véase el Apéndice 7.10), la distribución de probabilidad del cociente, bajo la hipótesis de que no hay ningún outlier, es conocida, y por lo tanto se puede calcular cuál es la probabilidad de obtener un valor como 0.3351 o superior sin que haya ningún outlier. En este caso, esta probabilidad es aproximadamente del 27%. Esta es la probabilidad de equivocarnos si declaramos la distancia de **a**1 como un outlier en el conjunto de las otras distancias. Es una probabilidad demasiado grande para arriesgarse en esta afirmación. Aunque la probabilidad de acertar es mayor (73%), no tomaremos ese riesgo.

Parece pues que **a**1 podría estar en el cluster 3. Pero hemos dicho que el cluster 1 no vamos a partirlo. ¿Qué hay de las demás letras del cluster 1 entonces? ¿Podríamos también integrarlas en el cluster 3? **a**4, **a**9 y **a**20 se encuentran a distancias de 0.3173, 0.2087, 0.2228 del modelo **a**8 del cluster 3. Usando el test de Dixon se encuentra que las tres caen por debajo del 5% de probabilidad de equivocación, si las declaramos ajenas al cluster 3 (1.56% para **a**9, 0.95% para **a**20 y menos de 0.01% para **a**4). Por tanto, lo hacemos así, y mantenemos definitivamente separados los clusters 1 y 3.

Repitiendo la idea con los demás clusters, se obtiene que:

· El cluster 5 se integra en el cluster 6 (de hecho, por muy poco no los hemos fusionado de entrada, al trazar la línea horizontal en la Fig. 8).

· El cluster 2 puede perfectamente integrarse en el cluster 1. En este caso, las dos distancias al modelo **a**5 son incluso menores que las de alguna letra ya incluida en el cluster 1.

· El cluster 2 también podría integrarse en el cluster 3, más forzadamente, ya que la probabilidad para la letra **a**19 es de 8.60%, de modo que con el 10% como probabilidad de equivocación permisible ya no lo integraríamos.

· El cluster 4 (letra **a**11) no se integra en el nuevo cluster unión del 5 y el 6 (letras **a**10, **a**12, **a**13), con probabilidad de equivocación p = 4.85%. Tampoco puede integrarse en los clusters unión de 1 y 2 (con p = 1.93%), ni en el cluster 3 (p = 3.32%). La letra **a**11 se queda por tanto sola definitivamente.

· Ya no hay ninguna posibilidad más de fusionar los nuevos clusters entre sí.

En definitiva, llegamos a la clasificación final de la tabla 48.

	cluster	modelo	media de las distancias	mínima, máxima	max/min
A	{4, 9, 1, 20, 18, 6, 15, 3, 5, 7, 19}	5	0.1114	0.0769, 0.1468	1.9090
B	{14, 21, 8, 16, 2, 17}	8	0.1137	0.0972, 0.1347	1.3858
C	{11}	11	–	–	–
D	{10, 12, 13}	12	0.1254	0.1225, 0.1281	1.0457

Tabla 48. Clasificación final en clusters de las letras a.

Remarquemos una vez más que el procedimiento que hemos usado es conservador: *estamos bastante seguros de que hay más de un cluster (4 al menos) en las 21 letras a, pero es perfectamente posible que haya más*.

Notas

5. "*Póliza. Se da este nombre al conjunto de letras, cifras, signos de puntuación, blancos, etc. que son necesarios para la composición de un determinado idioma. La póliza española difiere bastante de las extranjeras*". Bauer, Federico. *Conocimientos fundamentales para el aprendiz cajista*. Biblioteca Gráfica Neufville. Barcelona, 1921. Tomo I. Pág. 15

6. El Fondo Tipográfico Bauer-Neufville se encuentra actualmente en el Departamento de Diseño e Imagen de la Facultad de Bellas Artes en la Universidad de Barcelona.

7. Clasificación de tipos en la Biblia de 42 líneas

7.7 Representación gráfica de los *clustering*

El *Multidimensional Scaling* (MDS) es una técnica gráfica que permite representar en dimensión 2 o dimensión 3 los objetos que se están clasificando. Cada objeto viene representado por un punto en el espacio de dimensión 2 o 3, para dar una impresión visual de la cercanía o lejanía entre los distintos objetos o clusters de objetos. Las distancias que se observan en el gráfico no son verdaderas; son la mejor aproximación posible a las disimilaridades si se quieren representar los datos en un plano o en el espacio.

Las coordenadas concretas de los puntos en el Multidimensional Scaling no tienen significado intrínseco; sólo se pretende representar las distancias relativas entre los objetos. En particular, cualquier rotación, simetría o traslación del gráfico da lugar a otro gráfico con el mismo significado exactamente. Puede asegurarse, eso sí, que si un objeto tiene disimilaridades d_1 y d_2 con otros dos objetos, y d_1 es más pequeña que d_2, entonces las correspondientes distancias en el gráfico $\hat{d}_1$ y $\hat{d}_2$ también cumplirán que $\hat{d}_1$ es más pequeña que $\hat{d}_2$.

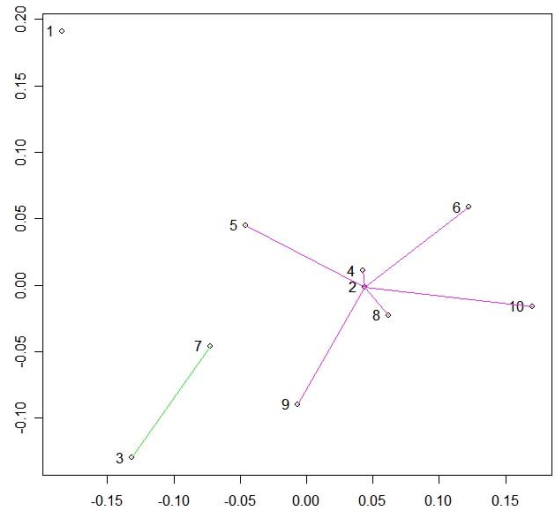


Fig. 137 MDS en dimensión 2 para la letra **i**.

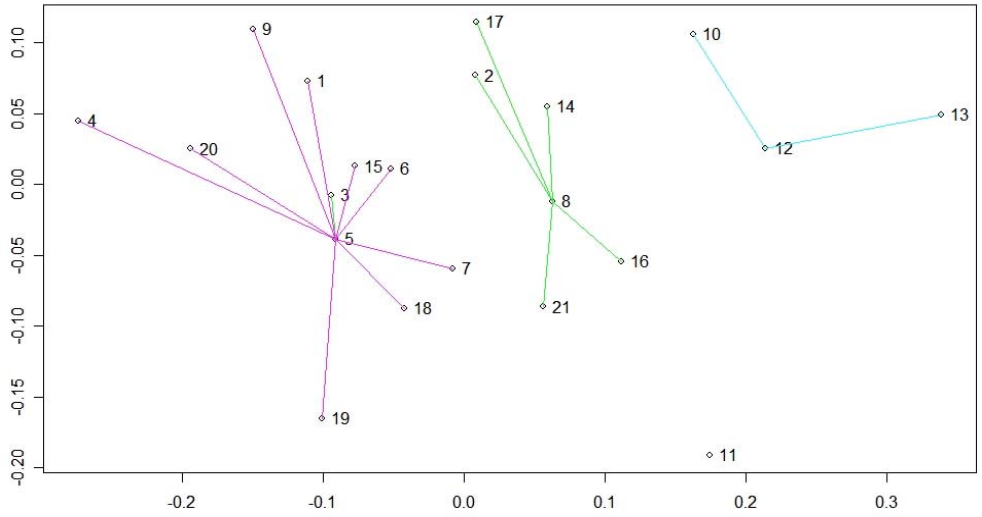
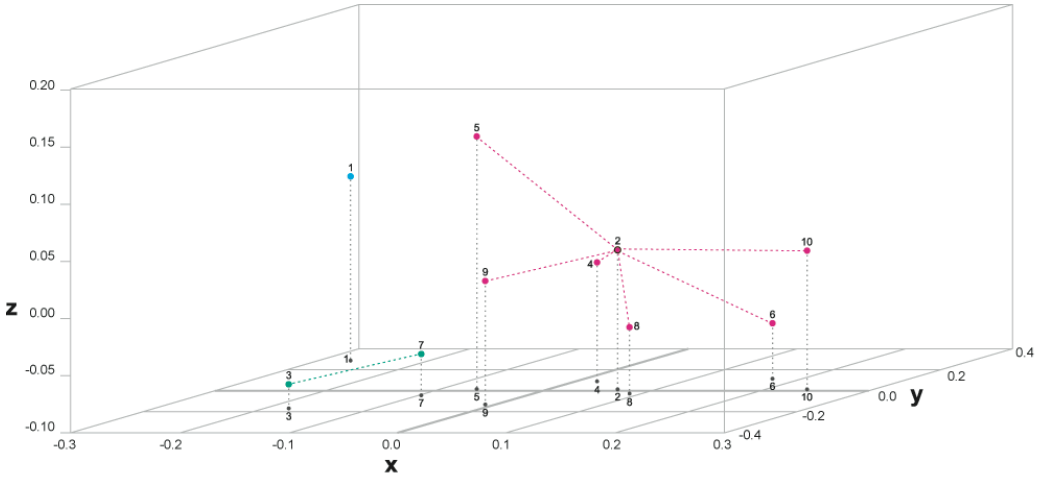


Fig. 138 MDS en dimensión 2 para la letra a.



	x	y	z
[1]	-0.254883160	0.2374716285	0.061301969
[2]	0.065679228	-0.0009140002	0.022255612
[3]	-0.172083938	-0.1823834378	-0.079060824
[4]	0.039612645	0.0359519044	0.004367450
[5]	-0.059363743	0.0035773565	0.120736716
[6]	0.162358946	0.1038255497	-0.051740631
[7]	-0.099351981	-0.0378659404	-0.063582022
[8]	0.093101932	-0.0262976573	-0.041759958
[9]	-0.009420086	-0.1416162486	0.007465108
[10]	0.234350158	0.0082508453	0.020016581

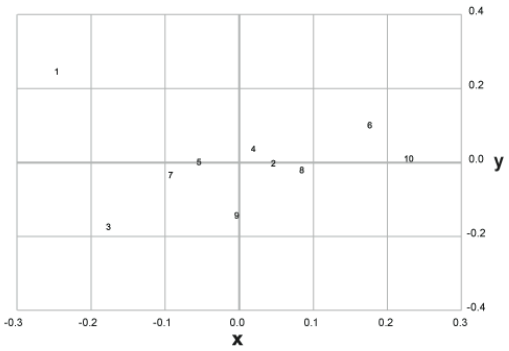
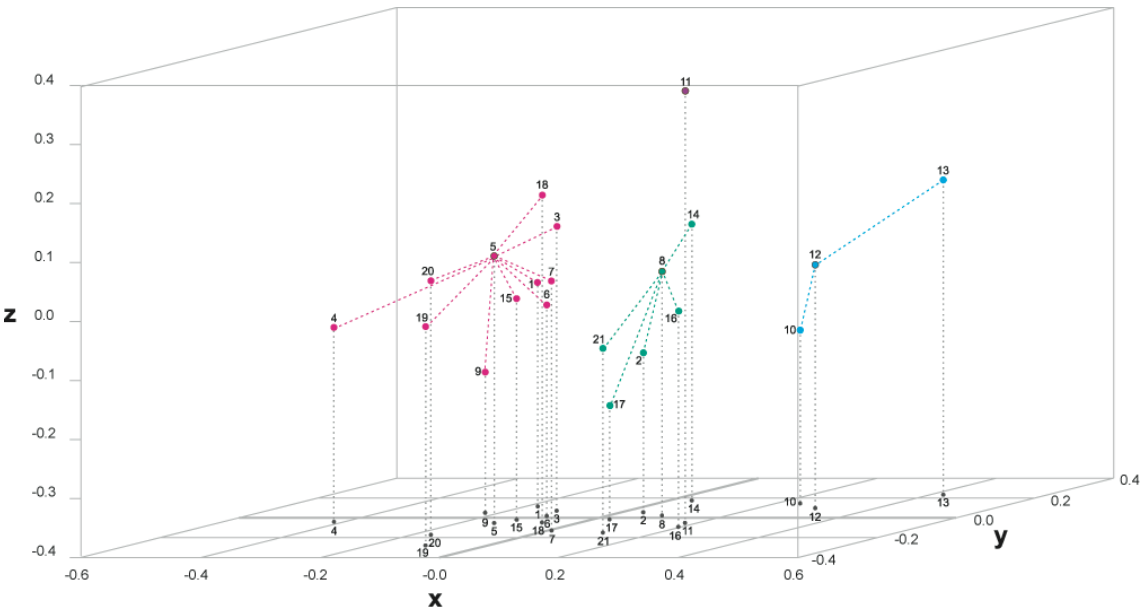


Fig. 139 MDS en dimensión 3 para la letra i.

Los MDS en dimensión 2 para las 10 letras i y las 21 letras a son los de las Figuras 137 y 138. Se han dibujado las estrellas que corresponden al clustering óptimo encontrado en cada caso. En las Figuras 139 y 140 se muestra el resultado de aplicar el MDS en dimensión 3.

En este caso para situar mejor visualmente las letras en su posición, se han listado sus coordenadas y se han trazado las proyecciones de cada punto sobre el plano horizontal inferior. Hay que tener en cuenta que estas proyecciones no tienen que ver con el MDS en dimensión 2: las posiciones de las letras en el espacio de dimensión 3 se calculan independientemente del resultado obtenido en dimensión 2.



	x	y	z
[1]	-0.17435511	0.117283728	-0.017133172
[2]	0.03029980	0.070027909	-0.129898506
[3]	-0.11225045	0.060017509	0.083390283
[4]	-0.40796945	-0.051045354	-0.065147809
[5]	-0.12778637	-0.070156983	0.057668733
[6]	-0.07892345	-0.001941919	-0.035674460
[7]	0.01089826	-0.135520928	0.022209570
[8]	0.10478298	0.003142395	0.016378414
[9]	-0.22291526	0.049956966	-0.159901774
[10]	0.22455688	0.167320016	-0.106082277
[11]	0.17654611	-0.048301469	0.334323002
[12]	0.31536493	0.072418810	0.015161355
[13]	0.43372871	0.217933972	0.136527330
[14]	0.04222312	0.168985733	0.068723730
[15]	-0.11713286	-0.028036171	-0.023333084
[16]	0.20106583	-0.097949901	-0.032624473
[17]	0.02785583	-0.009638567	-0.206506098
[18]	-0.05056607	-0.056312805	0.158104690
[19]	-0.09150214	-0.293433553	-0.026275278
[20]	-0.29736861	0.022318122	-0.002667287
[21]	0.11344730	-0.157067509	-0.087242887

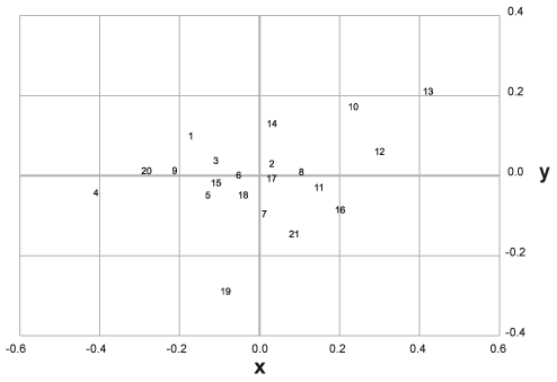


Fig. 140 MDS en dimensión 3 para la letra a.

7. Clasificación de tipos en la Biblia de 42 líneas

7.8 Espacios interiores

La letra a en esta tipografía posee dos espacios vacíos interiores, lo que da lugar a un contorno con tres curvas cerradas que no se tocan entre sí: exterior, interior superior e interior inferior. Es interesante estudiar alguna de las componentes interiores de las veintiuna a, clasificarlas como se ha hecho con las a completas, y comparar ambas clasificaciones.

Concretamente, hemos medido y clasificado los interiores inferiores con la misma técnica de los contornos completos. La existencia de diversos clusters en estos contornos interiores nos indicaría que *se usó más de un punzón en la manufactura de las matrices*.

A la vista del dendrograma de la Fig. 141, realizado con el método "average linkage", podemos conjeturar inmediatamente la existencia de cuatro clusters, formados por los contornos

$a = \{19, 21, 12, 2, 16\}$,
$b = \{8, 17, 10, 14, 11, 13\}$,
$c = \{4, 5\}$ y
$d = \{20, 1, 9, 3, 18, 7, 6, 15\}$.

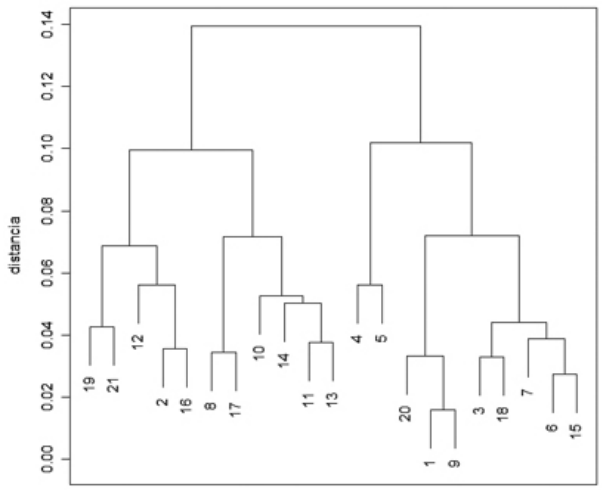
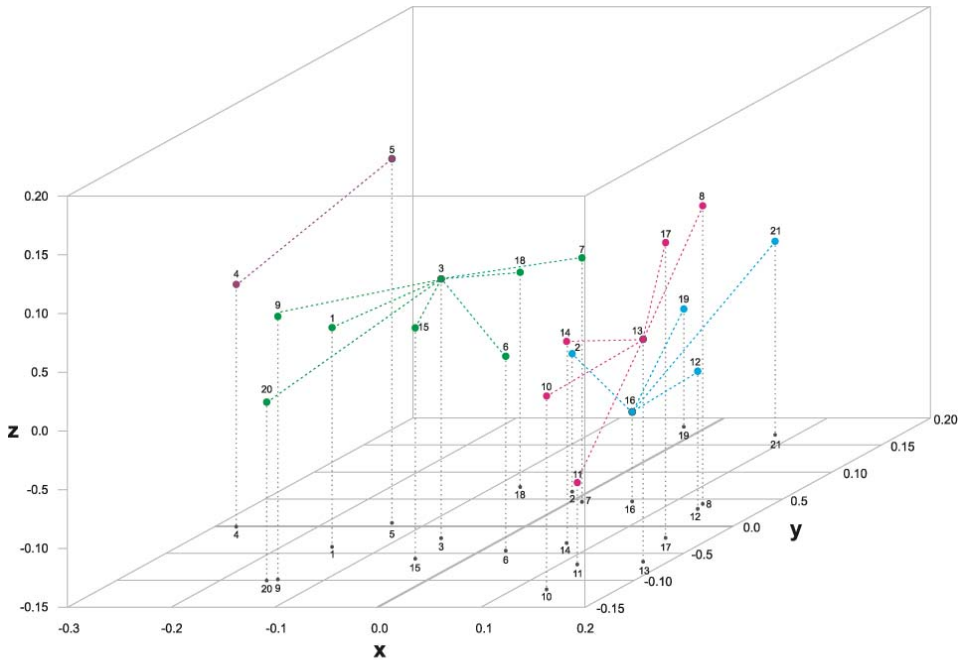


Fig. 141 Dendrograma para el interior inferior de la letra a, con el método "average linkage".

Buscaremos el contorno modelo de cada uno y aplicaremos los tests de outliers como en el Apartado 7.6.2 para confirmar que podemos mantenerlos separados con una probabilidad de equivocarnos muy pequeña.

Las letras modelo de los clusters anteriores resultan ser, respectivamente, a16, a13, a4 y a3. Efectuados los tests, se concluye que los cuatro clusters deben mantenerse separados. Si fijamos 0.05 como el límite máximo de probabilidad de equivocarnos, podemos afirmar que a4, a5, a20 no pertenecen al cluster *a*, que ninguna otra letra podría integrarse en el cluster *b*, y que a2, a4, a8, a1, a11, a12, a13, a14, a16, a17, a19, a21 no pueden integrarse en el *d*. Véase en el Apartado 7.10.5.1 los resultados numéricos.

La representación MDS de estos contornos en dimensión 3 es la de la Fig. 142.



	x	y	z
[1]	-0.15330518	-0.037109711	0.008384813
[2]	-0.01572098	0.062719687	-0.066409037
[3]	-0.06045095	-0.022163888	0.049516839
[4]	-0.28050580	-0.001860330	0.031003202
[5]	-0.13759835	0.006781649	0.148851839
[6]	0.02240057	-0.043942647	-0.014442780
[7]	0.01482543	0.041246693	0.035493521
[8]	0.13681405	0.035881595	0.089029094
[9]	-0.14955970	-0.095561851	0.051465161
[10]	0.13002847	-0.114325049	-0.014826360
[11]	0.12169475	-0.070596077	-0.120447012
[12]	0.13245791	0.040835400	-0.072525227
[13]	0.17511484	-0.063038687	0.012007039
[14]	0.07076047	-0.031054753	-0.006174385
[15]	-0.05108876	-0.058457032	0.020925736
[16]	0.06080675	0.044664468	-0.112626833
[17]	0.15606417	-0.020590079	0.083894947
[18]	-0.07513671	0.072850915	0.007882866
[19]	-0.02167330	0.183803226	-0.087569908
[20]	-0.15729521	-0.099309469	-0.028809662
[21]	0.08136754	0.169225939	-0.014623852

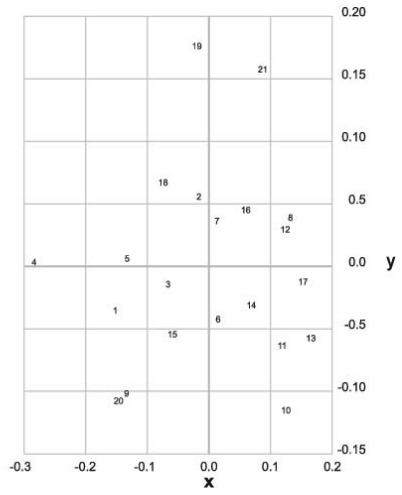


Fig. 142 MDS para el interior inferior de la letra a.

Por otra parte, podemos cruzar las clasificaciones de las a completas con las del contorno interior. A partir del MDS en dimensión 2 de la Fig. 10, en la que se representan las posiciones de los contornos completos de las a, hemos señalado los clusters A, B, C, D obtenidos de ellos con línea continua, y hemos superpuesto en líneas discontinuas los clusters a, b, c, d de los contornos interiores. Véase todo ello en la Fig. 143. En la Tabla 49 mostramos la misma información en otra forma.

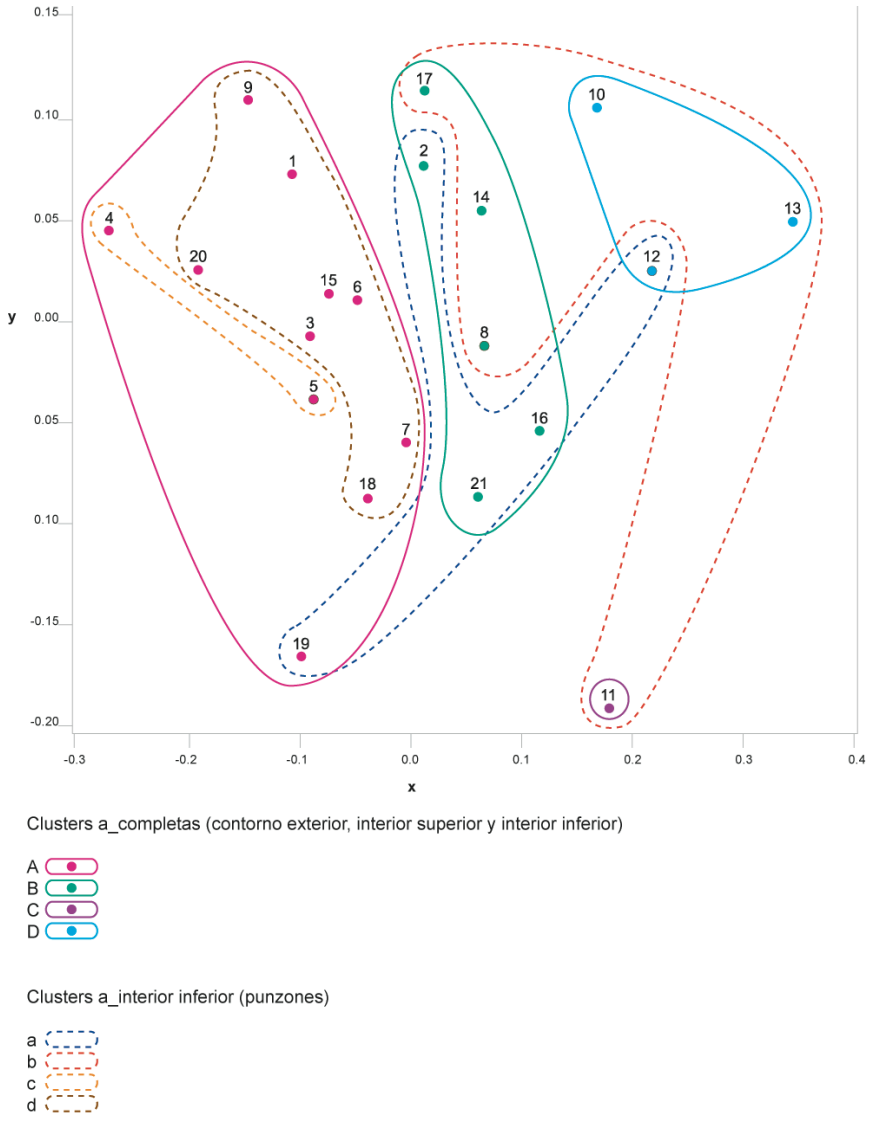


Fig. 143 MDS para las a completas con los clusters de los contornos interiores superpuestos.

Cluster	Punzón a	Punzón b	Punzón c	Punzón d
A	19		4, 5	1, 3, 6, 7, 9, 15, 18, 20
B	2, 16, 21	8, 14, 17		
C		11		
D	12	10, 13		

Tabla 49. Tabla cruzada entre los clusters de las a completas y los de sus contornos interiores inferiores.

En la tabla se observa una cierta (aunque no total) independencia entre la forma global de la letra y la forma de su centro inferior. Concretamente:

1. Letras cuyos contornos globales han sido clasificados como del mismo cluster, poseen centros inferiores clasificados en distintos clusters (*A* separado en *a, c, d*; *B* y *D* separados en *a, b*).
2. Letras cuyos contornos interiores han sido clasificados como del mismo cluster, poseen contornos globales clasificados en distintos clusters (*a* separado en *A, B, D*; *b* separado en *B, C, D*).

La explicación más razonable para este hecho es que los contornos exteriores, a diferencia de lo que hasta el momento se cree, también se realizaron con la misma técnica que los interiores es decir que *existió un punzón para crear el contorno exterior*.

Esto no es difícil de admitir si se recuerda que la gestación de la tipografía en metal proviene de la tradición de la orfebrería; más en concreto, de la acuñación de monedas – tradición que Gutenberg conocía a la perfección y que se ha explicado en el Capítulo 1 – por lo que sería razonable admitir que se tallaron diversos contornos exteriores que luego fueron trabajados según el caso con los mismos punzones para producir sus espacios interiores.

No se puede descartar, sin embargo, la posibilidad que nuestra clasificación de los contornos exteriores sea en realidad demasiado conservadora, y *pudieran haber más matrices distintas* involucradas en la página estudiada. Si, por ejemplo, subdividimos los clusters *A, B, D* según la Tabla 50, entonces no se sigue necesariamente la existencia de punzones para el contorno exterior, ya que a cada contorno global le corresponde un único punzón interior. En este momento nuestros datos son insuficientes para distinguir entre las dos posibilidades y este punto debería ser sin duda objeto de investigación ulterior.

Cluster	Punzón a	Punzón b	Punzón c	Punzón d
A1	19			
A2			4, 5	
A3				1, 3, 6, 7, 9, 15, 18, 20
B1	2, 16, 21			
B2		8, 14, 17		
C		11		
D1	12			
D2		10, 13		

Tabla 50. Tabla cruzada entre los clusters de las a completas y los de sus contornos interiores inferiores. (Ampliada).

Sin embargo, la alternativa representada en la Tabla 50 implica *la existencia de ocho matrices diferentes*, por lo que evidencia todavía más la conclusión principal de este estudio: ***existe una póliza múltiple en la impresión de la Biblia de 42 líneas.***

7. Clasificación de tipos en la Biblia de 42 líneas

7.9 Letras en páginas distintas

Como complemento al estudio realizado hasta aquí, se han medido algunas letras a de las páginas subsiguientes, escogidas según la pauta siguiente: Del volumen II, se han tomado las páginas 190 a 206 (correspondientes a todo el Evangelio de Mateo, incluido su prólogo) y siempre la primera columna del verso. Se ha tomado una letra de cada página: la primera que estuviera suficientemente aislada para poder ser medida sin error.

En total se midieron 17 letras, que denotaremos por a22–a38. No se pretende hacer con ellas una clasificación como la que se hizo con las letras a1–a22, sino solamente investigar a qué clusters de los ya construidos podrían razonablemente incorporarse. Para ello, compararemos cada una de las nuevas letras con los modelos de cada cluster de letras antiguas. La disimilaridad obtenida respecto al modelo se comparará con las disimilaridades de los otras letras del mismo cluster aplicando el test de Dixon para outliers, de la misma manera que se hizo en el Apartado 7.6.2 para confirmar que dos clusters eran realmente distintos.

Las mediciones se realizaron exactamente de la misma manera que con las primeras a, según se describe en el apartado 7.3, y resultó la tabla de disimilaridades que aparece en el Apéndice 7.10.3.

La conclusión más relevante que se ha obtenido es que tres de las nuevas letras no pueden integrarse (con una probabilidad máxima de error del 5%) con los clusters de las primeras veintiuna letras (Fig. 144). Se trata de a25, a31 y a33, que, como puede verse fácilmente en la tabla del Apéndice 7.10.3, mantienen una distancia comparativamente grande con los cuatro modelos. *Se deduce pues que había posiblemente más de cuatro matrices de letras* a *involucradas en la composición del Evangelio de Mateo*.

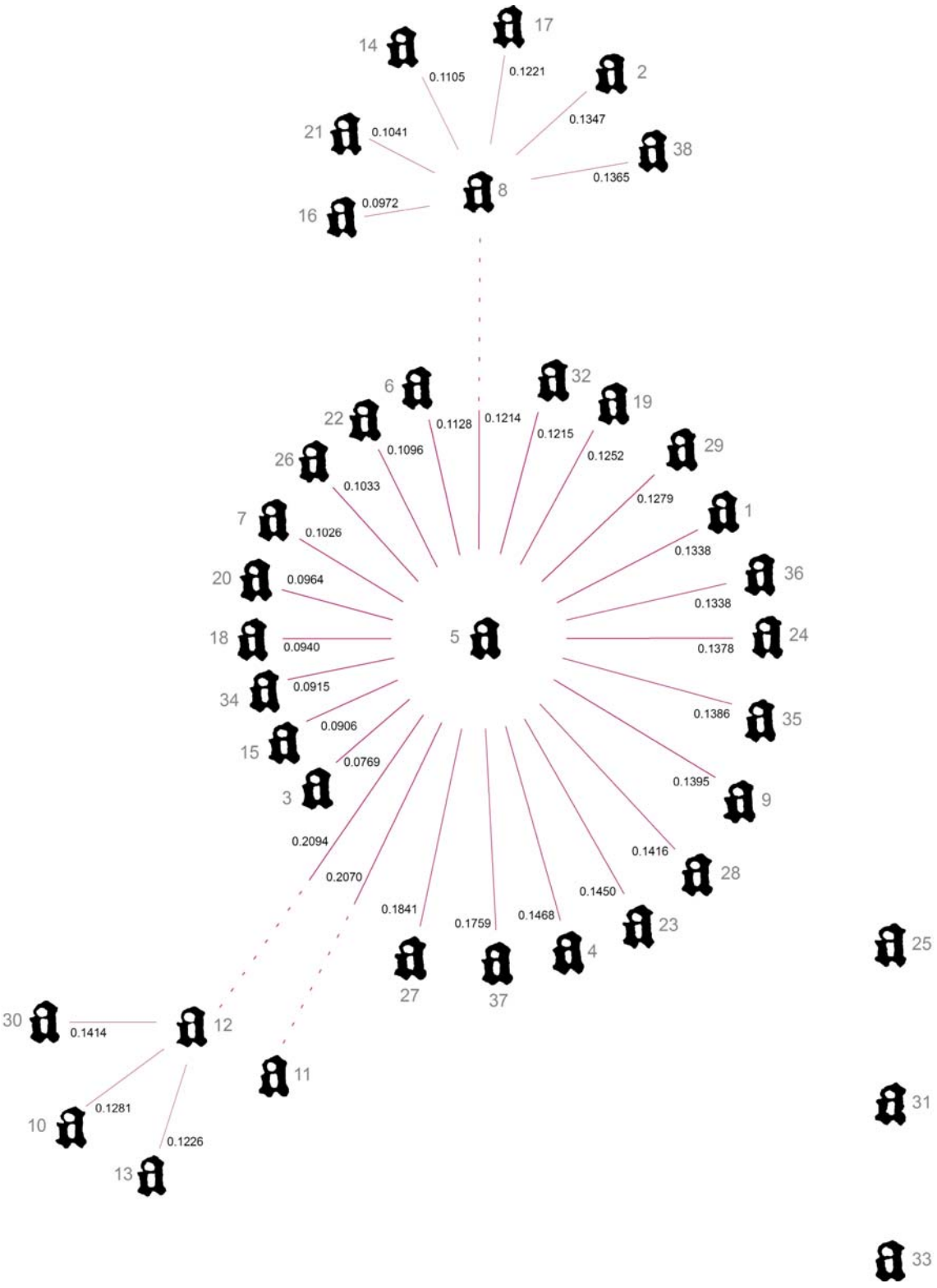


Fig. 144 Clustering integrando todas las letras a1 – a38.

7. Clasificación de tipos en la Biblia de 42 líneas

7.10 Apéndice

7.10.1 Mejores *clusterings* para i con dos y tres clusters

En el Apartado 7.4.3 se mencionan los clusterings óptimos con dos y tres clusters para la letra i, de entre los 511 y 9330 posibles, respectivamente. En las Tablas 51 y 52 se listan los veinte mejores clusterings de cada caso.

Orden	Valor	Modelos/distancia	Clustering
1	0.088129	(mods= 1 4, dist= 0.231040)	1 * 2 3 4 5 6 7 8 9 10
2	0.089731	(mods= 3 2, dist= 0.209735)	3 7 * 1 2 4 5 6 8 9 10
3	0.090777	(mods= 7 2, dist= 0.113851)	3 7 9 * 1 2 4 5 6 8 10
4	0.092746	(mods= 3 2, dist= 0.209735)	3 * 1 2 4 5 6 7 8 9 10
5	0.094429	(mods= 7 2, dist= 0.113851)	3 5 7 * 1 2 4 6 8 9 10
6	0.094905	(mods= 7 2, dist= 0.113851)	3 5 7 9 * 1 2 4 6 8 10
7	0.097599	(mods= 7 2, dist= 0.113851)	3 7 8 * 1 2 4 5 6 9 10
8	0.099970	(mods= 7 2, dist= 0.113851)	3 4 7 * 1 2 5 6 8 9 10
9	0.100428	(mods= 7 2, dist= 0.113851)	3 7 8 9 * 1 2 4 5 6 10
10	0.101204	(mods= 9 4, dist= 0.102566)	3 8 9 * 1 2 4 5 6 7 10
11	0.101210	(mods= 7 2, dist= 0.113851)	3 6 7 9 * 1 2 4 5 8 10
12	0.101480	(mods= 7 8, dist= 0.097677)	3 4 5 7 * 1 2 6 8 9 10
13	0.103092	(mods= 7 2, dist= 0.113851)	3 5 7 8 * 1 2 4 6 9 10
14	0.103223	(mods= 7 8, dist= 0.097677)	2 3 5 7 * 1 4 6 8 9 10
15	0.103273	(mods= 7 2, dist= 0.113851)	3 4 7 9 * 1 2 5 6 8 10
16	0.103552	(mods= 9 4, dist= 0.102566)	3 5 8 9 * 1 2 4 6 7 10
17	0.103785	(mods= 7 2, dist= 0.113851)	3 6 7 8 * 1 2 4 5 9 10
18	0.103906	(mods= 2 8, dist= 0.042524)	1 2 4 5 9 * 3 6 7 8 10
19	0.103937	(mods= 7 2, dist= 0.113851)	3 4 6 7 * 1 2 5 8 9 10
20	0.104188	(mods= 9 4, dist= 0.102566)	9 * 1 2 3 4 5 6 7 8 10

Tabla 51. Los veinte mejores clusterings con dos clusters para la letra i.

Las columnas de la tabla representan: El orden, la valoración del clustering, las letras modelo de cada cluster y distancia entre ellas, y los clusters (separados por asteriscos).

La correspondiente tabla cuando se pide que los clusterings contengan tres clusters es la siguiente:

Orden	Valor	Modelos	Clustering
1	0.066971	(mods= 1 3 2)	1 * 3 7 * 2 4 5 6 8 9 10
2	0.072722	(mods= 1 3 2)	1 * 3 * 2 4 5 6 7 8 9 10
3	0.078463	(mods= 1 7 2)	1 * 3 7 9 * 2 4 5 6 8 10
4	0.078602	(mods= 1 7 2)	1 * 3 4 7 * 2 5 6 8 9 10
5	0.082324	(mods= 1 7 2)	1 * 3 7 8 * 2 4 5 6 9 10
6	0.082386	(mods= 1 7 2)	1 * 3 4 7 9 * 2 5 6 8 10
7	0.083033	(mods= 1 5 8)	1 * 5 * 2 3 4 6 7 8 9 10
8	0.084868	(mods= 1 7 2)	1 * 3 7 8 9 * 2 4 5 6 10
9	0.084960	(mods= 1 7 2)	1 * 3 4 7 8 * 2 5 6 9 10
10	0.086066	(mods= 1 9 4)	1 * 9 * 2 3 4 5 6 7 8 10
11	0.086209	(mods= 1 2 7)	1 * 2 5 6 10 * 3 4 7 8 9
12	0.086420	(mods= 1 7 2)	1 * 7 * 2 3 4 5 6 8 9 10
13	0.087488	(mods= 1 10 4)	1 * 10 * 2 3 4 5 6 7 8 9
14	0.089461	(mods= 1 4 8)	1 * 4 5 * 2 3 6 7 8 9 10
15	0.089473	(mods= 1 6 4)	1 * 6 10 * 2 3 4 5 7 8 9
16	0.089577	(mods= 10 3 2)	10 * 3 7 * 1 2 4 5 6 8 9
17	0.089731	(mods= 3 7 2)	3 * 7 * 1 2 4 5 6 8 9 10
18	0.089784	(mods= 1 2 8)	1 * 2 5 * 3 4 6 7 8 9 10
19	0.089846	(mods= 1 4 8)	1 * 4 5 7 * 2 3 6 8 9 10
20	0.089922	(mods= 1 6 4)	1 * 6 * 2 3 4 5 7 8 9 10

Tabla 52. Los veinte mejores clusterings con tres clusters para la letra i.

7.10.2 Tabla de disimilaridades de las a completas

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20	a21
a1		0.1277	0.1165	0.1892	0.1338	0.1196	0.1301	0.1536	0.1049	0.1741	0.2653	0.2509	0.3712	0.1489	0.1315	0.2345	0.1866	0.1375	0.1939	0.0868	0.2262
a2	0.1277		0.1160	0.2370	0.1543	0.1391	0.1173	0.1347	0.1617	0.1643	0.2845	0.1522	0.2817	0.1422	0.1464	0.1169	0.1114	0.1723	0.1861	0.1834	0.1606
a3	0.1165	0.1160		0.1668	0.0769	0.1119	0.1474	0.1353	0.1471	0.2447	0.2214	0.2015	0.3191	0.0990	0.0897	0.1919	0.1715	0.1186	0.1720	0.0958	0.1771
a4	0.1892	0.2370	0.1668		0.1468	0.1637	0.2363	0.3173	0.1133	0.3779	0.3533	0.4261	0.5766	0.2478	0.1402	0.3280	0.2332	0.2330	0.2460	0.1057	0.3028
a5	0.1338	0.1543	0.0769	0.1468		0.1128	0.1026	0.1214	0.1395	0.2114	0.2070	0.2094	0.3265	0.1582	0.0906	0.1677	0.1749	0.0940	0.1252	0.0964	0.1563
a6	0.1196	0.1391	0.1119	0.1637	0.1128		0.1256	0.1274	0.1184	0.1778	0.2264	0.2161	0.3179	0.1294	0.0754	0.1651	0.1132	0.0938	0.1834	0.1421	0.1254
a7	0.1301	0.1173	0.1474	0.2363	0.1026	0.1256		0.1131	0.1669	0.1856	0.1928	0.2188	0.3023	0.1735	0.1425	0.1223	0.1489	0.1190	0.1161	0.1835	0.1296
a8	0.1536	0.1347	0.1353	0.3173	0.1214	0.1274	0.1131		0.2087	0.1327	0.1907	0.1179	0.2089	0.1105	0.1108	0.0972	0.1221	0.1289	0.1721	0.2228	0.1041
a9	0.1049	0.1617	0.1471	0.1133	0.1395	0.1184	0.1669	0.2087		0.2454	0.3359	0.3063	0.4550	0.1769	0.1224	0.2264	0.1415	0.1986	0.2144	0.1320	0.2119
a10	0.1741	0.1643	0.2447	0.3779	0.2114	0.1778	0.1856	0.1327	0.2454		0.2371	0.1281	0.1843	0.1964	0.2129	0.1809	0.1558	0.1694	0.3145	0.3417	0.2009
a11	0.2653	0.2845	0.2214	0.3533	0.2070	0.2264	0.1928	0.1907	0.3359	0.2371		0.2189	0.2366	0.1738	0.2451	0.1890	0.2468	0.1972	0.2894	0.3544	0.2449
a12	0.2509	0.1522	0.2015	0.4261	0.2094	0.2161	0.2188	0.1179	0.3063	0.1281	0.2189		0.1226	0.1593	0.2342	0.1809	0.2077	0.2282	0.2843	0.3288	0.1815
a13	0.3712	0.2817	0.3191	0.5766	0.3265	0.3179	0.3023	0.2089	0.4550	0.1843	0.2366	0.1226		0.1923	0.2185	0.1295	0.2995	0.3113	0.5016	0.5017	0.2912
a14	0.1489	0.1422	0.0990	0.2478	0.1582	0.1294	0.1735	0.1105	0.1769	0.1964	0.1738	0.1593	0.1923		0.1444	0.1752	0.1514	0.1907	0.2673	0.1665	0.1432
a15	0.1315	0.1464	0.0897	0.1402	0.0906	0.0754	0.1425	0.1108	0.1224	0.2129	0.2451	0.2342	0.3397	0.1444		0.1586	0.1224	0.1159	0.1655	0.1307	0.1629
a16	0.2345	0.1169	0.1919	0.3280	0.1677	0.1651	0.1223	0.0972	0.2264	0.1809	0.1890	0.1295	0.2185	0.1752	0.1586		0.1392	0.1809	0.1737	0.2762	0.1106
a17	0.1866	0.1114	0.1715	0.2332	0.1749	0.1132	0.1489	0.1221	0.1415	0.1558	0.2468	0.2077	0.2995	0.1514	0.1224	0.1392		0.1881	0.2208	0.1880	0.1372
a18	0.1375	0.1723	0.1186	0.2330	0.0940	0.0938	0.1190	0.1289	0.1986	0.1694	0.1972	0.2282	0.3113	0.1907	0.1159	0.1809	0.1881		0.1668	0.1922	0.1545
a19	0.1939	0.1861	0.1720	0.2460	0.1252	0.1834	0.1161	0.1721	0.2144	0.3145	0.2894	0.2843	0.5016	0.2673	0.1655	0.1737	0.2208	0.1668		0.1756	0.1438
a20	0.0868	0.1834	0.0958	0.1057	0.0964	0.1421	0.1835	0.2228	0.1320	0.3417	0.3544	0.3288	0.5017	0.1665	0.1307	0.2762	0.1880	0.1922	0.1756		0.2233
a21	0.2262	0.1606	0.1771	0.3028	0.1563	0.1254	0.1296	0.1041	0.2119	0.2009	0.2449	0.1815	0.2912	0.1432	0.1629	0.1106	0.1372	0.1545	0.1438	0.2233	

Tabla 53. Tabla de disimilaridades de las a (completas).

7.10.3 Tabla de disimilaridades a22-a38 respecto de los modelos establecidos

	5	8	11	12
22	0.1096	0.1096	0.2625	0.1993
23	0.1450	0.1907	0.3285	0.3097
24	0.1378	0.1609	0.2909	0.2914
25	0.2064	0.2183	0.4037	0.3658
26	0.1033	0.1237	0.2078	0.1517
27	0.1841	0.2499	0.3995	0.4447
28	0.1416	0.1598	0.2456	0.2392
29	0.1279	0.1395	0.2466	0.2525
30	0.1561	0.1671	0.2375	0.1414
31	0.2582	0.2313	0.3859	0.3686
32	0.1215	0.1572	0.2339	0.2426
33	0.2199	0.2209	0.2653	0.3190
34	0.0915	0.1605	0.2343	0.2560
35	0.1386	0.1519	0.2571	0.2295
36	0.1338	0.1413	0.1486	0.1929
37	0.1759	0.3131	0.3719	0.4254
38	0.1603	0.1365	0.2075	0.2268

Tabla 54. Disimilaridades entre las letras a22-a38 y las letras modelo.

7.10.4 Sobre el *Multidimensional Scaling*

La medida en que una posible representación gráfica de un conjunto de objetos difiere de las disimilaridades reales se denomina *stress* y puede definirse de varias maneras. La representación que hemos usado aquí es la variante conocida como *Kruskal Non-metric Multidimensional Scaling*. En ella, el stress S se define como

$$S = \sqrt{\frac{(f(d_{ij}) \quad \hat{d}_{ij})^2}{\hat{d}_{ij}^2}}$$

donde d son las disimilaridades dadas entre los objetos, $\hat{d}$ son las distancias en el gráfico, y f es una transformación monótona creciente que da una flexibilidad extra al ajuste. El gráfico se obtiene buscando las distancias $\hat{d}$ y la transformación f que hacen mínimo el stress. La monotonía de f asegura que si una disimilaridad es más pequeña que otra, entonces las distancias correspondientes en el gráfico respetan ese orden.

Más detalles sobre el Kruskal Non-metric Multidimensional Scaling pueden verse, por ejemplo, en el Capítulo 3 de [7], o en el Capítulo 11 de [8].

7.10.5 Sobre el test de outliers

La aplicabilidad del test de Dixon para la detección de outliers depende de unas ciertas condiciones, que en nuestro caso se cumplen. Éstas son:

a) Los datos siguen una ley gaussiana
b) son estadísticamente independientes

La gaussianidad se justifica por el hecho de que cada uno de los datos es la media de una gran cantidad de valores pequeños. Por lo tanto, independientemente de la ley de probabilidad que siga cada uno de esos pequeños valores, su media será aproximadamente gaussiana[7].

La independencia es evidente puesto que el resultado de la medición de una distancia entre una letra cualquiera y una letra "modelo" no influye en el resultado de la medición de la distancia entre una tercera letra y el mismo modelo. (Si incluyéramos la distancia entre las letras que no son modelo, la independencia se perdería.)

En ocasiones, un outlier queda enmascarado por la presencia de otro dato dudoso. Existen tests estadísticos para poner de manifiesto dos outliers a la vez (como el *test de Grubbs*, véase por ejemplo [3]); sin embargo, siguiendo nuestro espíritu conservador, hemos preferido dejar que pudiera existir en algún caso este enmascaramiento, y aceptar ambos datos como genuinos.

Puede parecer por contra más arriesgada la decisión de declarar separados dos clusters aún cuando sólo una de las letras de un clúster es rechazada como perteneciente al otro cluster. Hay que tener en cuenta, por un lado, que es todavía más arriesgado partir uno de los clusters que se ha decidido no dividir, y por otra parte, que la probabilidad de equivocarse al declarar fuera de un cluster más de una letra es más pequeña que la probabilidad de declarar fuera de un cluster a cada una de ellas. Por lo tanto, si una de las letras no supera la probabilidad de equivocación fijada del 5%, un conjunto de más de una letra tampoco lo superará.

El resto de este apartado esta dedicado a mostrar los *p-valores* (probabilidades de error al declarar una letra ajena a un cluster) para cada una de las letras y clusters con un *p*-valor más pequeño que 0.05. Los *p*-valores están redondeados a cuatro decimales.

Letras a completas

Tests de Dixon para la clasificación en clusters de las letras a completas. Estos son los resultados que hemos usado para obtener, a partir de la Tabla 47, la clasificación definitiva de la Tabla 48.

Respecto del Cluster 1 (letra modelo: 5):

Letra	Cluster	p-valor
10	5	0.0294
11	4	0.0368
12	6	0.0326
13	6	0.0000

Respecto del Cluster 3 (letra modelo: 8):

Letra	Cluster	p-valor
4	1	0.0000
9	1	0.0156
11	4	0.0332
13	6	0.0155
20	1	0.0095

Respecto de los Clusters 1 y 2 unidos (letra modelo: 5):

Letra	Cluster	p-valor
10	5	0.0146
11	4	0.0193
12	6	0.0167
13	6	0.0000

Respecto de los Clusters 5 y 6 unidos (letra modelo: 12):

Letra	Cluster	p-valor
1	1	0.0360
6	1	0.0500
7	2	0.0485
9	1	0.0250
11	4	0.0485
15	1	0.0416
18	1	0.0441
19	2	0.0284
20	1	0.0223

Contornos interiores de las letras a

Para los clusters de los contornos interiores, se partió de cuatro clusters (véase Apartado 7.8) y, en base al test de Dixon, se decidió mantenerlos separados. A continuación se muestran los p-valores menores de 0.05 que se obtuvieron. Nótese que respecto del Cluster C no tiene sentido aplicar el test, puesto que éste sólo contiene una distancia.

Respecto del Cluster A (letra modelo: 16) :

Letra	Cluster	p-valor
4	c	0.0057
5	c	0.0302
20	d	0.0314

Respecto del Cluster B (letra modelo: 13):

Letra	Cluster	p-valor
1	d	0.0000
2	a	0.0000
3	d	0.0000
4	c	0.0000
5	c	0.0000
6	d	0.0472
7	d	0.0039
9	d	0.0000
12	a	0.0180
15	d	0.0030
16	a	0.0038
18	d	0.0000
19	a	0.0000
20	d	0.0000
21	a	0.0000

Respecto del Cluster D (letra modelo: 3):

Letra	Cluster	p-valor
2	a	0.0468
4	c	0.0000
8	b	0.0000
10	b	0.0000
11	b	0.0000
12	a	0.0000
13	b	0.0000
14	b	0.0484
16	a	0.0000
17	b	0.0000
19	a	0.0000
21	a	0.0000

Letras i

Para las diez letras i, obtuvimos en el Apartado 7.4 los clustering óptimos (exactos), condicionando a la existencia de dos y de tres clusters. Podemos también aplicar el test de Dixon para outliers a esos clusterings, y el resultado nos confirma las subdivisiones obtenidas. De paso, podemos usar el test para ver si una subdivisión en 4 clusters sería también razonable. El resultado será negativo, como vamos a ver.

El mejor clustering en cuatro clusters (exacto, con el método del Apartado 7.4) es

$$\alpha = \{1\}\ ,\ \beta = \{3\}\ ,\ \gamma = \{5\}\ ,\ \delta = \{2, 4, 6, 7, 8, 9, 10\}\ .$$

Pero el test de Dixon no sostiene i_5 fuera del cluster δ. En cambio sí lo hace para i_1 e i_3:

Respecto del Cluster (letra modelo: 8):

Letra	Cluster	p-valor
1	α	0.0000
3	β	0.0117

El segundo mejor clustering con cuatro clusters corresponde a

$$\alpha = \{1\}\ ,\ \beta = \{3\}\ ,\ \gamma = \{7\}\ ,\ \delta = \{2, 4, 5, 6, 8, 9, 10\}\ .$$

que parece más consistente con nuestra división anterior en tres clusters. Otra vez i_1 e i_3 no se integran en δ,

Respecto del Cluster (letra modelo: 2):

Letra	Cluster	p-valor
1	α	0.0032
3	β	0.0075

y, aunque i_7 sí lo haría, antes lo integraríamos con i_3, a la vista del dendrograma, y dado que no podemos hacer el test de outliers contra un cluster de un sólo elemento.

Como curiosidad, obsérvese que las letras modelos del cluster δ en los dos casos anteriores son distintas. De hecho, i_2 e i_8 son letras muy cercanas y no son sorprendentes estos cambios al variar ligeramente la composición de un cluster.

En conclusión, este estudio forzando cuatro clusters no nos aporta ya más que lo que observábamos con dos y tres clusters: Podemos postular con bastante seguridad la existencia de tres clusters, y en cuanto a su composición concreta, la apuesta más segura es nuestra primera opción

$$\alpha = \{1\}\ ,\ \beta = \{3,\ 7\}\ ,\ \delta = \{2,\ 4,\ 5,\ 6,\ 8,\ 9,\ 10\}\ ,$$

con un cierto riesgo, acotado por una probabilidad del 32% de equivocarnos al separar la letra 7 del cluster δ. El clustering de dos clusters, uniendo β y δ, nos parece extremadamente conservador.

Referencias

[1] Jerome Friedman, Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning* , Springer, 2001.
[2] A. D. Gordon. *Classification*, Chapman and Hall / CRC, 1999.
[3] V. Barnett, T. Lewis. *Outliers in statistical data*, Wiley, 1984.
[4] The R Project for Statistical Computing, http:\\www.r-project.org.
[5] S-Plus, http://www.insightful.com.
[6] Mitutoyo, http://www.mitutoyo.com.
[7] T.F. Cox, M.A.A. Cox. *Multidimensional scaling* (2nd ed.), Chapman and Hall / CRC, 2001.
[8] W. N. Venables, B.D. Ripley. *Modern applied statistics with S* (4th ed.), Springer, 2002.

Notas

1. Técnicamente, el término "distancia", en el sentido matemático habitual, debe cumplir la desigualdad triangular: la distancia entre dos objetos ha de ser menor que la suma de las distancias de esos objetos a un tercero. Esto no lo cumple la medida de disimilaridad que vamos a usar.
2. Gutenberg, Johannes. *Biblia Latina*. Moguntiae, (c. 1454- agosto, 1456). Alemania. Depositada en el Fondo Antiguo de la Biblioteca de la Universidad de Sevilla.
3. El centro de gravedad de un conjunto de puntos con coordenadas (x_1, y_1), ..., (x_n, y_n) es el punto

$$\frac{1}{n}(x_1 + \ldots + x_n\ ), \frac{1}{n}(y_1 + \ldots + y_n\ )$$

4. Véase la nota 1.
5. "*Póliza. Se da este nombre al conjunto de letras, cifras, signos de puntuación, blancos, etc. que son necesarios para la composición de un determinado idioma. La póliza española difiere bastante de las extranjeras*". Bauer, Federico. *Conocimientos fundamentales para el aprendiz cajista*. Biblioteca Gráfica Neufville. Barcelona, 1921. Tomo I. Pág. 15
6. El Fondo Tipográfico Bauer-Neufville se encuentra actualmente en el Departamento de Diseño e Imagen de la Facultad de Bellas Artes en la Universidad de Barcelona.
7. Consecuencia del resultado conocido en la Teoría de la Probabilidad como *Teorema Central del Límite*.