Análisis de Correspondencias de Matrices de Abundancia Arqueológicas

Resumen

En este capítulo se exponen las técnicas de Análisis de Correspondencias (CA) usadas en otros capítulos de este libro. CA es presentado como una técnica de análisis exploratorio multivariado, como una técnica de análisis de proximidades basado en las distancias de Benzécri, como una técnica para descomponer la chi-cuadrado total de matrices de frecuencias, y como un método de mínimos cuadrados para ajustar modelos de asociación o de ordenación.

1 Introducción

El *Análisis de Correspondencias* (CA a partir de aquí) es una técnica para analizar matrices de datos numéricos no negativos. CA está relacionado con el *análisis de componentes principales (PCA)* y el *escalamiento multidimensional (MDS),* es decir, es una forma de *análisis de proximidades*. CA es aplicado más frecuentemente a tablas rectangulares de frecuencias, también conocidas como *tabulaciones cruzadas* o *tablas de contingencia,* aunque las aplicaciones a matrices de binarias o de presencia-ausencia son también bastante comunes.

Esta técnica estadística es usada más a menudo para analizar tabulaciones cruzadas, computando y evaluando alguna medida de *independencia* o *homogeneidad,* tal como Chi-cuadrado. En el análisis de la independencia investigamos si el cuerpo de la tabla es el producto de los marginales. O, si uno prefiere una formulación asimétrica, si las filas de la tabla difieren sólo porque tienen diferentes totales de fila (y las columnas sólo difieren porque tienen diferentes totales de columna).

La Chi-Cuadrado de Pearson y otras medidas relacionadas cuantifican como es de diferente la tabla observada respecto de la tabla esperada, calculada a partir de los totales de fila y columna. Los residuales de Pearson son usados para investigar las desviaciones respecto de la independencia. CA complementa este análisis Chi-cuadrado clásico, puesto que hace tanto una *descomposición* y una *representación gráfica* de las desviaciones respecto de la independencia.

1.1 Historia

CA tiene una historia complicada, tanto en estadística como en arqueología. La prehistoria de CA, comenzando con el trabajo realizado por Pearson alrededor de 1900 y acabando con la reinvención de la técnica hecha por Fisher y Guttman alrededor de 1940, es expuesta en De Leeuw (1983). Posteriormente la técnica fue reinventada bajo muchos nombres diferentes, en muchos países diferentes, y en muchas disciplinas científicas. Nuevas reencarnaciones todavía continúan apareciendo, aunque a un ritmo más lento que antes, en la literatura sobre minería de datos y análisis de datos. Beh (2004) es una revisión bibliográfica completa reciente.

La historia de CA en arqueología es expuesta por Baxter (1994, p. 133-139). Aunque hubo algunas aplicaciones previas a ejemplos arqueológicos en la literatura sobre CA, el mérito de la introducción de la técnica a los arqueólogos normalmente es atribuido a Bølviken y otros (1982). Las aplicaciones anteriores casi sin excepción vinieron de arqueólogos de la Europa continental, bajo la influencia, sin duda, de la escuela francesa de *Analyse des Données,* bajo el liderazgo de Benzécri (1973a, 1973b). Una buena revisión de estas aplicaciones arqueológicas continentales del CA es la de, por ejemplo, Müller y Zimmerman [1997].

A partir de la exposición de Baxter es claro que los arqueólogos en la Europa continental iban por delante de los arqueólogos en Gran Bretaña, los cuales se subieron a bordo alrededor de 1990. Clive Orton, uno de los decanos de la arqueología cuantitativa en Gran Bretaña, planteó que CA era la técnica más importante introducida en arqueología en los años 80 (Orton, 1999, p.32). El CA arqueológico migró de Gran

Bretaña a los Estados Unidos donde llegó poco antes del 2000. Duff (1996, p. 90) indicó, en un artículo muy influyente de mitad de los 90, que el CA "no estaba bien instalado en la literatura Americanista". Y, muy recientemente, Smith y Neiman concluyeron: "CA tiene una larga historia de uso por parte de los arqueólogs en Europa continental pero su uso por los arqueólogos Americanistas es a la vez más reciente y más raro." (Smith y Neiman, 2007, p.55).

Hay varias razones posibles por las que el CA no se convirtió rápidamente en popular en arqueología en Gran Bretaña y en Estados Unidos. La más importante, quizás, es que los metodólogos arqueólogos tienden a buscar orientación entre los especialistas en estadística, y en la estadística el CA no fue realmente conocido hasta 1980, a pesar del trabajo de Hill (1974). Excepto in Francia, naturalmente, pero la estadística francesa estaba aislada relativamente de la estadística predominante. Las técnicas multivariantes dominantes aplicadas en arqueología eran el MDS y el PCA (algunas veces disfrazado como análisis factorial). El trabajo más influyente en el área en los años 70 fue el de Hodson y otros (1971), los cuales se concentraron en las técnicas de MDS de Boneva, Kendall y Kruskal. Todas estas técnicas son formas de análisis de proximidad, pero todas ellas difieren del CA de diversas maneras.

LeBlanc (1975, p. 22) predijo, en un artículo pionero: "El análisis de proximidades parece contener una gran cantidad de promesas y en el futuro suplantará con toda probabilidad otros métodos de seriación." Si interpretamos esta predicción de modo estricto, en términos de los métodos que estaban disponibles ya en 1975, ello fue incorrecto, por razones que son bastante obvias en retrospectiva. Los datos, en arqueología y en cualquier otro lado, vienen en formas muy diferentes. Algunas veces tratamos con tabulaciones cruzadas, otras con matrices de incidencia, y en otras con datos multivariados que describen objetos arqueológicos en términos de un número de variables cualitativas o cuantitativas. No hay ninguna razón para esperar que una técnica que está diseñada para un tipo particular de datos también funcionará, o ni siquiera será la apropiada, para otro tipo de datos. Una técnica de análisis de datos debe obviamente tener en cuenta la naturaleza de los datos, y obligar a todos los datos a un formato común de "proximidades" puede no ser la estrategia óptima. Pero las ventajas básicas del análisis de proximidades mencionadas por LeBlanc [1975, p. 22] son todavía acertadas. "En el pasado, el objetivo básico de la seriación ha sido ordenar una serie de unidades culturales sobre el supuesto de una única variable subyacente, normalmente el tiempo. Ahora es posible seriar unidades según dos o más variables usando alguna forma de análisis de proximidad o MDS. Esto aumenta el poder de la seriación en gran medida, y entre otras ventajas, da una idea mucho mejor de la adecuación de los datos a una variable (por ejemplo, solo el tiempo) que la proporcionada por los métodos anteriores".

Puesto que el CA fue redescubierto y reintroducido en diferentes países en momentos diferentes, la mayoría de los autores arqueológos se sienten obligados a dar algún tipo de introducción a la técnica. Esto es cierto incluso para los artículos recientes tales como Poblome y Groenen [2003] y Smith y Neiman [2007]. Nuestro análisis del CA difiere en algunos aspectos de los que tradicionalmente se encuentran en la arqueología. En otros aspectos es bastante estándar. En primer lugar, y esto es bastante común, no se presenta la técnica exclusivamente como un método de seriación. Puede haber muchas razones diferentes por las que los yacimientos arqueológicos son similares o diferentes y, para citar a Kruskal [1971], "El tiempo no es la única dimensión." La mayoría de las

gráficos de CA son, por supuesto, mapas bidimensionales en el plano, lo cual ya sugiere que más de una dimensión puede ser relevante. En segundo lugar, hablamos de CA, tanto como técnica exploratoria y como método de ajustar un modelo estadístico concreto. Y, finalmente, relacionamos el ajuste mínimo cuadrático del CA con el ajuste máximo verosímil del modelo de la Distancia Exponencial (ED). Tanto ED como CA ordinario pueden ser considerados formas alternativas, estrechamente relacionadas, de análisis de correspondencias.

1.2. Tipos y atributos

LeBlanc [1975] compara la *seriación de tipo* y la *seriación de atributo*. Véase también Duff [1996]. Se puede explicar esta comparación distinguiendo entre los diferentes tipos de datos a los que el CA se puede aplicar. En el contexto del CA, la seriación de atributo corresponde al análisis de correspondencias múltiples (MCA), tratado en Gifi [1990, capítulo 3], y la seriación de tipo corresponde al CA simple, tratado en Gifi [1990, Capítulo 8]. O bien, para traducir esto en términos de software, la seriación de atributo se corresponde con el paquete homals en R [De Leeuw y Mair, 2008a], mientras que la seriación de tipo se corresponde con el paquete anacor [De Leeuw y Mair, 2008b].

LeBlanc [1975, p. 24] distingue cuidadosamente los términos "atributo", "tipo", "variable", y "dimensión". Realmente, él utiliza "variable" y "dimensión" intercambiablemente, pero es probablemente una buena idea reservar "dimensión" para los ejes en las representaciones multidimensionales de los datos. Una "variable" es entonces un aspecto definido formalmente del grupo de objetos en el estudio. Cada variable es medida en términos de una escala, y las características mutuamente excluyentes de la escala son llamadas "atributos". En el libro de Gifi [1990], una variable se define de manera similar como una proyección de los objetos en el estudio dentro de las categorías de una variable. Definir un número de variables sobre un conjunto de objetos crea, en la terminología del sistema de software de R [R Development Core Team, 2007], un “marco de datos" (dataframe). Más específico para la arqueología es la noción de un "tipo", que Leblanc define como "la existencia de una asociación no aleatoria entre los atributos de dos o más dimensiones" [LeBlanc, 1975, p. 24]. Así, los tipos son agregaciones de atributos en diferentes variables, y por consiguiente se pueden recontar más fácilmente, y son más susceptibles de ser tratados con técnicas basadas en frecuencias.

Esta exposición hace también posible comparar el CA con el MDS y con el PCA. En el MDS el primer paso es generalmente derivar algún tipo de matriz simétrica de similaridades entre yacimientos, ensamblajes, orígenes, o unidades culturales. Hay muchas maneras de definir las similaridades, y en muchos casos, la elección de una medida de similaridad particular es algo arbitrario. Además, en lugar de calcular similaridades entre yacimientos, también podríamos decidir calcular similaridades entre variables que describen los artefactos encontrados en los yacimientos. Una medida de similaridad entre las variables de uso común es el coeficiente de correlación. No está claro cómo se relaciona el análisis MDS de los yacimientos y el análisis MDS de las variables. En el PCA se suele empezar con una matriz de correlación entre las variables, y luego derivar cargas de componentes para describir las variables y las puntuaciones de los componentes para describir los yacimientos. Esto significa que el PCA puede ser

usado para hacer gráficos del conjunto, también conocido como biplot [Gower y Hand, 1996]. Los biplots permiten visualizar la información multidimensional de una manera muy atractiva, y como tal, van más allá de la simple seriación.

Una desventaja del PCA mencionada a menudo es que asume que las relaciones entre las variables son lineales. Esto, sin embargo, ya no es cierto para versiones modernas no lineales de PCA, revisadas por ejemplo, en De Leeuw [2006]. Además, hay una estrecha relación entre el PCA no lineal y MCA, tan cerca que, de hecho, el PCA no lineal puede llevarse a cabo con el paquete MCA `homals` [De Leeuw y Mair, 2008a].

El marco del análisis de correspondencia de Gifi [1990] ofrece una sola clase de técnicas para analizar los atributos de matrices artefactos por variables, matrices de frecuencias de tipos por yacimientos, y matrices de incidencia de tipos por yacimientos. Se trata básicamente de, para usar un término del Analyse des données de Benzécri, de una cuestión de "codage"(codificación). Uno puede codificar tanto tipos y yacimientos como atributos de artefactos, y entonces la tabla de frecuencias de tipo por yacimientos consiste simplemente en la tabulación cruzada de esas dos variables.

Una ventaja importante del CA y el MCA sobre el MDS y el PCA es que están lo más cerca posible de los datos originales, no importa si los datos son frecuencias o incidencias o variables con atributos. No hay necesidad de en primer lugar, elegir una medida de similaridad o de correlación, y no hay necesidad de agregar datos en matrices de correlación o productos. Es verdad que el CA puede ser presentado en términos de una determinada medida de desigualdad, la distancia de Benzécri. Nosotros daremos tal presentación en este artículo. Pero es sólo una interpretación de la técnica, y las distancias de Benzécri tienen una estrecha conexión con las chi-cuadrado habituales que pueden ser calculadas a partir de las frecuencias.

1.3 Aplicaciones Típicas en Arqueología

Discutiremos algunas de las aplicaciones típicas del CA en arqueología con más detalle, para ilustrar dónde la técnica puede ser apropiada y en lo que los arqueólogos se fijan.

En Bølviken et al. [1982], se usaron tres conjuntos de datos de la Edad de Piedra en el norte de Noruega. El primero, proveniente de Iversfjord, utiliza treinta y siete tipos líticos en catorce “house site assemblages”. Debido a dificultades de interpretación el análisis fue repetido después de agrupar los treinta y siete tipos en nueve categorías de herramientas. El gráfico conjunto en dos dimensiones de las casa y las categorías de herramientas es interpretado en términos de orientación económica y permanencia del asentamiento. El segundo ejemplo es para la Edad de Piedra Temprana en la zona del fiordo de Varanger. Los datos son los recuentos de frecuencias de 16 tipos funcionales de herramientas en 43 yacimientos. Los gráficos de dos dimensiones dan un refinamiento que es interpretado en términos de hipótesis arqueológicas cualitativas anteriores. El análisis fue repetido agrupando las herramientas en siete clases, produciendo resultados menos informativos. En el tercer ejemplo CA fue utilizado para establecer una cronología. Los datos provenían de una granja en la isla de Helgøy en Troms. Hay diecinueve clases de objetos distribuidos en 15 capas de excavación, fechados por carbono entre los siglos catorce al diecinueve AC. El análisis muestra las capas proyectadas en una curva en herradura de dos dimensiones. Las proyecciones sobre la curva pueden ser utilizadas para reordenar las filas y columnas de la matriz de datos, produciendo una seriación que corresponde muy cercanamente a la basada en

datación por carbono.
El artículo de Duff [1996] sobre micro-seriación compara seriación de atributo y de tipo, siguiendo a LeBlanc [1975]. Pero mientras que LeBlanc utilizó escalamiento multidimensional para la seriación tipo, Duff utilizó CA. Los datos son recuentos de seis tipos cerámicos en 40 lugares en el Pueblo de Las Muertas, en la región de Zuni (Cibola) en Nuevo México, desde el siglo XIII hasta el siglo XIV. La solución CA de dos dimensiones muestra una herradura débil, con mucha dispersión a su alrededor, pero produce en esencia el mismo orden de las unidades producido por el análisis de MDS de Leblanc.

Una aplicación inicial de CA a los materiales Americanistas es Clouse [1999], que utilizó CA para analizar los artefactos encontrados en las excavaciones en el asentamiento militar en Fort Snelling, Minnesota. Los yacimientos son ocho edificios de defensa, once edificios de apoyo, y ocho edificios de habitaciones. En todos los yacimientos los artefactos fueron contados y clasificados en catorce grupos, tales como culinarios, armamento, comercio, mobiliario. Matrices de abundancia separadas son dadas para defensa, apoyo, y edificios de vivienda y CA diferentes fueron calculados por separado. Tanto los gráficos conjuntos, mostrando unidades y grupos de artefactos en dos dimensiones, y gráficos de unidad, que sólo muestran las unidades, fueron presentados. Las agrupaciones de las unidades se ajustan a lo que se espera sobre la base del Modelo de Yacimiento Militar, pero proporcionan información más detallada. Clouse [1999, p. 105] sostiene que CA produce que detalles, bien esperados o bien inusuales, sean más claramente visibles que el resumen numérico dado por la tabla.

El excelente artículo de Smith y Neiman [2007] se propone comparar seriación de frecuencias, en la tradición de Ford [1952], con el CA. Se estudian dos casos. En el primer caso estudian el área de la Costa del Golfo, cerca de los ríos Chattahoochee y Apalachicola, en Alabama, Georgia y Florida. Los datos proceden de los períodos Woodland Medio y Tardío (100 AC al 900 DC). Datos sobre cerámica fueron recolectados en muchos yacimientos, de los cuales 29 fueron seleccionados, por ser los que tenían más de 80 trozos pintados. Los 29 yacimientos fueron subdivididos en 84 “assemblages” y los trozos fueron clasificados en 18 tipos de alfarería. Obviamente, será importante para el resultado final de la técnica cómo los artefactos y los lugares de donde provienen son agrupados en filas y columnas de la tabla. El CA de los 84 “assemblages” muestra un patrón muy claro de herradura, con una agrupación clara de los yacimientos a lo largo de la curva. "Los resultados del CA confirman lo que la pura solución de seriación sugiere: no hay una fuente significativa de variación en las frecuencias aparte del tiempo." [Smith y Neiman, 2007, p. 61] El análisis fue repetido tras eliminar algunos de los conjuntos. Este CA más pequeño fue validado (como método de seriación) por medio de un gráfico de las puntuaciones CA frente a la datación por radiocarbón para yacimientos seleccionados.

El segundo estudio de caso en el artículo de Smith y Neyman proviene de Kolomoki, un sitio “multimount” bien documentado en el sudoeste de Georgia. Se trata de un análisis dentro de un solo yacimiento, no un análisis con varios yacimientos. El CA utiliza 20 “assemblages” y nueve tipos de cerámicas. Gráficos de dos dimensiones separados de “assemblages” y tipos no muestran un efecto de herradura, sino una segunda dimensión interpretable y significativa. La solución CA muestra efectos, por ejemplo los de tipo espacial, no detectables por la seriación de frecuencias de una dimensión. La primera

dimensión CA es validada otra vez como temporalidad, usando datos de radiocarbono. Usaremos el conjunto de datos de Kolomoki como uno de los ejemplos ilustrativos en este capítulo.

2 Seriación

Hay un interesante desarrollo histórico paralelo de lo que en términos generales se podría llamar "métodos de seriación" en psicometría, ecología y arqueología. Los principales pasos de estos desarrollos se producen en el mismo orden, pero en diferentes momentos en el tiempo, de una manera semejante a artefactos arqueológicos en yacimientos diferentes. Veamos primero psicometría.

2.1 Psicometría

En la década de 1940, en el departamento de la guerra, Guttman [1944] descubrió el análisis de escalogramas, un método para ordenar simultáneamente ítems de acierto/error o de actitudes (columnas) y al mismo tiempo respondentes (filas), con datos de una matriz de datos binarios. Inicialmente, las escalas fueron construidas por ensayo y error, de tal manera que que las filas y columnas de la matriz de datos binarios eran permutadas para crear la propiedad de “aquellos consecutivos”. Más precisamente, se buscaba ordenar filas y columnas de tal manera que todos los aquellos estén junto al otro. Esto era hecho manualmente, usando varios dispositivos ingeniosos. Al mismo tiempo, la teoría para los cálculos basados en componentes principales ya estaba disponible Guttman [1941, 1950]. De hecho, Guttman [1941] es el primer artículo que define rigurosamente MCA, y Guttman [1950] demuestra rigurosamente que la primera dimensión de MCA proporciona la ordenación consecutiva para datos libres de error. El monumental libro de Coombs [1964] hizo una presentación sistemática de estas técnicas de lápiz y papel, aplicadas a los diversos análisis de proximidades. Aunque el marco conceptual de Coombs sigue siendo pertinente, estas técnicas fueron superadas por métodos de cálculo por ordenador que ya estaban disponibles cuando el libro apareció.

2.2 Arqueología

Los métodos de Guttman fueron publicados alrededor de 1950, casi simultáneamente con Robinson [1951]. Para hablar sobre este trabajo, tomaremos prestada un poco de la terminología de Kendall [1969]. Una matriz de incidencias de, por ejemplo, yacimientos por tipos, es una *matriz de Petrie* o *P-matriz* si en cada columna todos los unos ocurren consecutivamente. Una matriz simétrica no-negativa es una *matriz de Robinson* o *R-matriz* si las filas y las columnas son unimodales y alcanzan sus valores máximos en la diagonal. Por unimodal queremos decir que las entradas aumentan a un máximo para luego disminuir de nuevo. Similaridades entre yacimientos cuya matriz de incidencia es una P-matriz a menudo forman una R-matriz. Una vez más, aquí hay una conexión interesante con la psicometría. En la definición original del modelo de Spearman para la inteligencia general, que se remonta a 1904, una batería de pruebas satisfacería el modelo si su matriz de correlación fuera una matriz-R.

La noción de matriz-P puede ser generalizada a las matrices de abundancia, es decir, a cualquier matriz con entradas no-negativas. Una matriz de abundancia es una *matriz-Q* si sus columnas son unimodales. Esto es lo mismo que decir que las columnas de la matriz de abundancia pueden ser representadas como una serie de gráficas de barcos

acorazados, similares a los de Ford [1952] o Smith y Neiman [2007]. Muchas de las técnicas de seriación arqueológica originales propuestas por Petrie, Robinson, Ford, Hole y Shaw, y otros toman una matriz de abundancias o de incidencias y permutan los yacimientos de tal manera que se convierten en una matriz-P o en una matriz-Q. La permutación que es encontrada entonces ordena los yacimientos en el tiempo, es decir, es una seriación. En última instancia, sin embargo, especialmente para grandes matrices, encontrar combinaciones óptimas es lo que se conoce en la ciencia de la computación, como NP-duro, lo cual básicamente significa que el problema de optimización, aunque finito, no puede ser resuelto en una cantidad de tiempo práctico, incluso utilizando los computadores más rápidos Arlif [1995].

Una forma de evitar que los cálculos involucrados con permutaciones sean impracticables es utilizar otras definiciones relacionadas de óptimo. Como hemos señalado anteriormente, Guttman probó ya en 1950 que el CA se puede utilizar para encontrar la permutación óptima a una matriz-P en el caso de no error. Para matrices de abundancia, véase también Gifi [1990, Capítulo 9], o Schriever [1983]. De hecho, estos artículos prueban más. También muestran que, en el caso de no error la segunda dimensión del CA será una función cuadrática de la primera, es decir, representar los sitios en el plano mostrará una curva cuadrática.

Kendall [1971] y otros posteriormente desarrollaron el bien conocido programa HORSHU que aplica el MDS a las similaridades derivadas de las matrices de abundancia, y entonces deriva el orden de la proyección de los yacimientos sobre la herradura o arco curvilínea. "Vemos el arco como un indicador relativamente benigno que los datos subyacentes, de hecho, contienen curvas con forma de barco acorazado." [Smith y Neiman, 2007, p. 60].

2.3 Ecología

En ecología el concepto clave es el de un "gradiente". El énfasis en el análisis de datos no está en el tiempo, como en la arqueología, sino en las características del medio ambiente. Lo que se llama "seriación" en la arqueología se denomina "ordenación" en ecología [Gauch, Jr., 1982]. Plantas y animales funcionan bien en determinadas circunstancias, y muy bien, por ejemplo, en un nivel óptimo de humedad o de altitud. Las diferentes especies necesitan diferentes altitudes y / o diferentes grados de humedad. En ecología, por supuesto, tenemos la gran ventaja que los gradientes ambientales como la altitud puede medirse directamente. Esto es a diferencia de la psicometría, donde la aptitud y la actitud son constructos teóricos, y a diferencia de la arqueología, en donde la información directa sobre el origen en el tiempo de un artefacto suele faltar. Así que la ecología tiene el Análisis de Gradientes Directo, en el que representamos las frecuencias de las especies como una función del gradiente. En muchos casos observamos distribuciones unimodales, es decir, la matriz de abundancia es una matriz-Q.

Inicialmente, igual que en psicometría y en arqueología, las técnicas de ordenación usaron métodos de lápiz y papel para reordenar las filas y columnas de la matriz de abundancia, o de las matrices de similaridades derivadas con una estructura de Robinson [Whittaker, 1978]. Esto cambió con la llegada del computador, y, como en arqueología y psicometría, los ecologistas pasaron al PCA y al MDS para hacer la ordenación, así como a una gran multitud de medidas de semejanza o similaridad.

El CA fue introducido en ecología por Hill [1974] como "promediado recíproco". Ter Braak [1985] mostró cómo el CA estaba relacionado con el modelo de respuesta unimodal, sin entrar en detalles matemáticos precisos. Los ecólogos inicialmente estaban preocupados por el efecto herradura, ya que lo consideraban un mero artefacto, sin ningún significado empírico. Ahora sabemos con más precisión de donde vienen las estructuras arqueadas, y sabemos que indican efectos unidimensionales muy fuertes. Véase, en particular, Schriever [1985], o van Rijckevorsel [1987]. Estamos, por consiguiente, contentos si vemos una herradura con claridad, especialmente en arqueología, donde tenemos aún quizás más razones para esperar la unidimensionalidad.

Discutiremos la relación entre los modelos de respuesta unimodal, en particular el modelo Gaussiano de Ihm y van Groenewoud [1975], con más detalle en la sección 7 sobre el modelo de Distancia Exponencial.

3 Matrices de Abundancia

Formalizaremos a continuación algunos de los conceptos que hemos mencionado en la introducción. Consideremos una tabla $r \times c$ N con *recuentos*. Las filas corresponden con los *yacimientos* r, las columnas con los *tipos* c. La frecuencia n_{ij} indica cuantas veces el tipo j fue hallado en el yacimiento i. Esta matriz N se llama una *matriz de abundancia*. También definimos la sumas por fila $n_{i\bullet}$ y las sumas por columna $n_{\bullet j}$ de la table. El *gran total* $n_{\bullet\bullet}$ es la suma de todos los recuentos en la tabla, el cual abreviaremos simplemente a n.

Debería quizás ser mencionado que las *matrices de presencia-ausencia* o *matrices de incidencia* son un caso especial de las matrices de abundancia, en las que todas las entradas de la tabla son bien cero o uno. Una entrada se limita a indicar si un tipo está presente en un yacimiento o no. Esto significa que nuestro análisis de las matrices de abundancia abarca también las matrices de presencia-ausencia.

Hay un tipo más general de matriz de datos, el cual es muy común también en arqueología. Supongamos que la unidad de observación es un artefacto tal y como un fragmento de cerámica, una pieza de obsidiana, o quizás un hueso de pescado. Las unidades pueden ser descritas en términos de una serie de variables que pueden ser bien cualitativas (categóricas) o cuantitativas (numéricas). La matriz de abundancia es un caso muy especial de esto, en la sólo que hay dos variables categóricas utilizadas para describir las unidades, es decir, *yacimiento* y *tipo*.

Los datos de abundancia N pueden ser codificados como una matriz $n \times 2$, donde n es el gran total de la tabla, y donde la primera columna es el yacimiento y la segunda el *tipo*. La tabla N es entonces la *tabulación cruzada*, o la *tabla de contingencia*, de las dos variables. Pero es evidente que en un caso general variables tales como tamaño, color, peso, composición podrían utilizarse también. Para estos datos multivariados más generales, necesitamos una técnica como el MCA, también conocido como *análisis de homogeneidad*, [Gifi, 1990, Greenacre y Blasius, 2006]. Dado que los datos analizados en este libro son todos del formato de tablas de contingencia bivariado más simple, no vamos a discutir MCA más allá de aquí. Como mencionamos en la introducción, es la técnica ideal para seriación basada en el atributo en el sentido de LeBlanc [1975], en el cual no se agregan nuestros datos a tipos y “assemblages”, y a recuentos en una tabla de

contingencia.

3.1 Ejemplos

A lo largo del capítulo vamos a usar dos ejemplos para ilustrar los conceptos del CA. El primer ejemplo de una matriz de abundancia proviene de una matriz más grande de recuentos de fragmentos por yacimientos por tipos de cerámica. Todas las muestras son de colecciones de superficie hechas hacia 1940 en Jalisco, México por Kelley [1945].

Este ejemplo no es una aplicación realista del CA porque es demasiado pequeño y demasiado simple. Los resultados del CA en realidad no añaden nada a lo que podemos ver fácilmente con sólo mirar la tabla, pero este mismo hecho hace que el ejemplo sea útil para ilustrar los conceptos básicos y los cálculos.

Insertar Tabla 1 por aquí

Los códigos para los tipos, utilizados como encabezados de columna, son

- AutPol: Autlan policromada (Autlan Polychrome);
- MiReBr: Diversos rojos sobre marrón, beige (Miscellaneous Red on Brown, Buff);
- AuWhRe: Autlan blanco sobre rojo (Autlan White on Red);
- AltRed: Cerámica Roja Altillos (Altillos Red Ware).

Los yacimientos son

- Yacimiento 21, Cofradía N º 1, y yacimiento 34, Hacienda Nueva, se incluyen en la Cofradía Complex (temprana);
- Yacimiento 23, Cofradía No. 3, y yacimiento 37, Amilpa, se incluyen en el Complejo Mylpa (intermedio);
- Yacimiento 7, Mezquitlan y yacimiento 9, Altillos, se incluyen en el Complejo Autlan (finales).

El segundo ejemplo son los datos de restos cerámicos de las sepultura Kolomoki en Georgia [Sears, 1956, Pluckhahn, 2003], analizados previamente con CA por Smith y Neiman [2007]. Ya hemos discutido estos datos en la introducción. Hay “assemblages” y 9 tipos de cerámica en los datos.

4 Matrices Asociadas

Con la matriz de abundancia podemos asociar algunas otras matrices. En primer lugar está la matriz *P* de *proporciones*, cuyos elementos se definen por

$$p_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}.$$

La matriz con las proporciones muestra más claramente cómo los recuentos se distribuyen entre las celdas. De nuevo, los marginales de fila son: $p_{i\bullet}$ y los marginales de columna son $p_{\bullet j}$.

Insertar Tabla 2 por aquí

4.1 Independencia

Decimos que la variable de fila (yacimiento) y la variable de columna (tipo) son *independientes* si $p_{ij}=p_{i\bullet}p_{\bullet j}$. La independencia puede interpretarse en el sentido de que el cuerpo de la tabla no da información adicional, de hecho, toda la información está contenida en los marginales. Si sabemos la frecuencia relativa de los yacimientos y los tipos, entonces podemos predecir exactamente cuántos de cada tipo habrá en cada yacimiento.

Medimos la independencia en CA mediante la llamada *inercia*, tomando prestado el término de la física. Definir la tabla Z de *residuales de Pearson* con

$$z_{ij} = \frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}}}.$$

Los elementos de Z muestran la desviación entre la proporción observada y la proporción esperada en la hipótesis de independencia (corregido el error típico de la proporción). Elementos positivos indican que vemos más en la celda correspondiente de lo que esperábamos, los elementos negativos significa que vemos menos. La inercia se define simplemente como

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} z_{ij}^2.$$

En el ejemplo de Kelley la inercia es 0,9338, y los residuales de Pearson están en la Tabla 3.

Insertar Tabla 3 por aquí

Si los datos son una muestra aleatoria, y si los tipos y los yacimientos son independientes, entonces nX^2 se distribuye como una variable aleatoria chi-cuadrado con $(r-1)(c-1)=15$ grados de libertad. En nuestro ejemplo nX^2 es 1207,508. Además cada uno de los $\sqrt{n}z_{ij}$ es aproximadamente normal típico, es decir, que son lo que se conoce comúnmente como una puntuación z, y pueden ser evaluados según su importancia en la forma habitual. Las puntuaciones z se encuentran en la Tabla 4.

Insertar Tabla 4 por aquí

Es evidente que en el ejemplo de Kelley la inercia total es demasiado grande, las puntuaciones z son en su mayoría enormemente significativas, y las dos variables *yacimiento* y *tipo* están muy lejos de ser independientes. Por supuesto, en la mayoría de aplicaciones arqueológicas los datos están muy lejos de ser muestras aleatorias, puesto que generalmente enumeramos y clasificamos todos los artefactos encontrados en el yacimiento. Sin embargo, todavía podemos tomar la inercia como una guía para indicar cuánta estructura hay en los datos, o, más exactamente, cuánta estructura hay en los datos que no puede ser predicha a partir de los marginales.

4.2 Condicionando en filas y columnas

En estudios arqueológicos la hipótesis de independencia no es la forma más natural de examinar matrices de abundancia. La independencia es el concepto apropiado si la tabla

de contingencia viene de una muestra aleatoria de una distribución bivariada discreta, es decir, si muestreamos tanto yacimientos y tipos. Normalmente, sin embargo, los yacimientos no han sido muestreados. Han sido fijados bien por diseño o bien por circunstancias geográficas.

Lo que nos interesa realmente es comparar la distribución de los tipos en los diferentes yacimientos que hemos seleccionado. Así, nos interesa principalmente comparar las filas de la matriz de abundancia, ya que cada fila define una distribución sobre tipos. Afortunadamente, la hipótesis de la homogeneidad de las filas es matemáticamente equivalente a la hipótesis de la independencia. Podemos ver esto más fácilmente normalizando las filas, dividiendo cada fila por su suma de fila.

Para mantener nuestro tratamiento simétrico, también consideramos el caso (menos común en la arqueología) en que puede ser interesante o conveniente comparar también las columnas. Usando las sumas de filas y las de columnas, podemos normalizar la tabla de frecuencias (o equivalentemente la tabla con las proporciones) dividiendo las entradas de la tabla por sus marginales de fila o de columna. Esto define dos nuevas tablas, la primera condicionada por las filas, la segunda condicionada por las columnas. Los elementos son definidos por

$$p_{j|i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{p_{ij}}{p_{i\bullet}},$$
$$p_{i|j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{p_{ij}}{p_{\bullet j}}.$$

La hipótesis de la independencia $p_{ij}=p_{i\bullet}p_{\bullet j}$ puede ahora ser escrita en dos formas equivalentes

$$p_{j|i} = p_{\bullet j},$$
$$p_{i|j} = p_{i\bullet},$$

que podemos llamar *homogeneidad de las filas* y *homogeneidad de las columnas*. La homogeneidad de filas dice que la distribución de probabilidad de los tipos es la misma para todos los yacimientos. La homogeneidad de las columnas dice que la distribución de probabilidad de los yacimientos es la misma para todos los tipos, que en nuestro contexto parece una forma menos natural de expresar el mismo hecho básico matemático.

La tabla 5 muestra la distribución de los tipos en cada uno de los yacimientos, y en la última fila la distribución de los tipos sobre todos los yacimientos, es decir, $p_{\bullet j}$. Tenemos homogeneidad si y sólo si todas las filas de la tabla, incluyendo la última fila, son las mismas. La tabla 6 muestra la distribución de los yacimientos en cada uno de los tipos, y en la última columna la distribución de los yacimientos en todos los tipos, es decir, la $p_{i\bullet}$. Tenemos homogeneidad si y sólo si todas las columnas de la tabla, incluyendo la última columna, son las mismas.

Podemos definir medidas apropiadas de homogeneidad de filas y columnas. Estas son de nuevo denominadas *inercias* en CA. Así que ahora hay una inercia para cada fila, y una para cada columna. Se definen por

$$X^2_{i\bullet} = \sum_{j=1}^{c} \frac{(p_{j|i} - p_{\bullet j})^2}{p_{\bullet j}},$$
$$X^2_{\bullet j} = \sum_{i=1}^{r} \frac{(p_{i|j} - p_{i\bullet})^2}{p_{i\bullet}}.$$

Las filas con una gran inercia difieren de la línea media, es decir, el vector $p_{\bullet j}$ de proporciones marginales por columna. Y las columnas con una gran inercia difieren de la columna de promedios $p_{i\bullet}$.

Anteriormente, hemos definido la *inercia total*. Debido a la relación simple

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{i\bullet}p_{\bullet j})^2}{p_{i\bullet}p_{\bullet j}} =$$
$$= \sum_{i=1}^{r} p_{i\bullet}X^2_{i\bullet} = \sum_{j=1}^{c} p_{\bullet j}X^2_{\bullet j}$$

la inercia total es la suma ponderada de las inercias de las filas y las columnas.

Insertar Tabla 5 por aquí

Insertar Tabla 6 por aquí

Bajo la hipótesis de un muestreo aleatorio de yacimientos y homogeneidad de filas, la $nX^2_{i\bullet}$ se distribuyen según chi-cuadrado con $c-1$ grados de libertad. Si tenemos muestreo aleatorio y homogeneidad de columnas, la $nX^2_{\bullet j}$ se distribuye según chi-cuadrado con $r-1$ grados de libertad.

5 Análisis de Correspondencias exploratorio

El propósito básico del CA exploratorio es hacer un *mapa de los tipos* y un *mapa de los yacimientos*. Al decir "mapa" nos referimos a una representación geométrica con un número reducido de dimensiones. Si elegimos dimensionalidad igual a dos, por ejemplo, un mapa de los tipos se compone de c puntos en el plano, con un punto correspondiente a cada tipo. Si elegimos dimensionalidad tres, entonces un mapa de los yacimientos se compone de r puntos en el espacio tridimensional. A veces, un mapa unidimensional, el cual pone todos los yacimientos en línea recta, ya es suficiente para presentar la información esencial en la tabla.

La ubicación de los puntos en el mapa no es arbitraria, por supuesto. Si hacemos un mapa bidimensional de los tipos, por ejemplo, queremos que las distancias entre los puntos c, en el plano sean aproximadamente iguales a las distancias entre las columnas c de la matriz de abundancia N. Y lo mismo para el mapa de los yacimientos y las filas de N.

Distancia en el mapa se define de la manera habitual como "en línea recta (as the crow flies)". En otras palabras, es una distancia euclídea ordinaria. Pero la distancia entre las

columnas de la matriz de abundancia usa pesos que tienen en cuenta la estabilidad estadística de los recuentos de células. En concreto, en CA usamos *distancias de Benzécri* (también conocidas como *distancias chi-cuadrado*). La distancia de *Benzécri* al cuadrado entre la fila *i* y la fila *k* de la tabla *N* viene dada por

$$\delta_{ik}^2 = \sum_{j=1}^{m} \frac{(p_{jli} - p_{jlk})^2}{p_{\bullet j}},$$

y la distancia al cuadrado de *Benzécri* entre la columna *j* y la columna *l* de la tabla *N* es

$$\delta_{jl}^2 = \sum_{i=1}^{n} \frac{(p_{il\,j} - p_{il\,l})^2}{p_{i\bullet}}.$$

En las tablas 7 y 8 damos las distancias al cuadrado de Benzécri para las filas y las columnas en el ejemplo de Kelley.

Insertar Tabla 7 por aquí

Insertar Tabla 8 por aquí

Si miramos más de cerca a la tabla 7 ya podemos predecir lo que hará el CA. Si queremos una representación geométrica en la que las distancias aproximen las distancias de Benzécri, entonces está bien claro cómo se vería tal representación. Las distancias de Benzécri entre los yacimientos 21 y 34 y entre los yacimientos 23 y 37 son casi cero. Así, en un mapa, los yacimientos 21 y 34 van a coincidir, y los yacimientos 23 y 37 también coincidirán. Los yacimientos 9 y 7 están cerca también, y (21,34) es aproximadamente igual de distante que los dos grupos (7,9) y (23,37). Un mapa de dos dimensiones aparecerá como un triángulo isósceles con los tres grupos de yacimientos en las esquinas. El lado más corto esta en algún lugar alrededor de $\sqrt{2}$ or $\sqrt{3}$, los dos lados más largos están alrededor de $\sqrt{6}$.. También vemos que en general será imposible asignar la información de distancia en una línea recta, porque en ese caso tendríamos que permitir que (7,9) coincidiesen con (23,37). En este pequeño ejemplo podemos fácilmente ver como se vería un mapa, pero en ejemplos más grandes, como el de Kolomoki, esto se vuelve mucho más complicado. Por eso tenemos CA, el cual aproxima las distancias Benzécri por medio de distancias euclídeas en el mapa de una manera precisa.

En CA aproximamos las distancias Benzecri *por abajo*. Vamos a explicar este concepto. En cualquier mapa de CA de los yacimientos, por ejemplo, siempre tendremos $d_{ik} \leq \delta_{ij}$ij, donde d_{ik} es la distancia euclídea entre los puntos *i* y *k* en el mapa. Más precisamente, el CA genera una secuencia de mapas, el primero tiene una sola dimensión, el segundo tiene dos, y así sucesivamente. El mapa final tiene $t=\min(r-1,c-1)$ dimensiones, es decir, 3 en el ejemplo de Kelley y 8 en el ejemplo de Kolomoki. Los mapas están *anidados*, en el sentido de que la proyección sobre la primera dimensión de todos los mapas es idéntica a la de un mapa unidimensional y la proyección sobre el plano de las dos primeras dimensiones para todos los mapas con dimensión de al menos dos es igual al mapa de dos dimensiones. Y así sucesivamente. Si $d_{ik}^{(s)}$ son las distancias en mapa *s*-dimensional, con $1 \leq s \leq t$, entonces

$$d_{ik}^{(1)} \leq d_{ik}^{(2)} \leq \ldots \leq d_{ik}^{(t)} = \delta_{ik}.$$

Así, el mapa t-dimensional tiene distancias exactamente iguales a las distancias de Benzécri. Mapas en menos dimensiones aproximan las distancias, y la aproximación mejora, para cada una de las distancias, cuando aumenta la dimensionalidad. La aproximación es desde abajo, porque las distancias del mapa son siempre menores que las distancias Benzécri, no importa cuál sea la dimensionalidad del mapa. Por supuesto, el mismo razonamiento se aplica a las distancias de Benzécri distancias entre las columnas y el mapa del CA para los tipos.

El mapa no sólo aproxima las distancias de Benzécri entre los yacimientos o tipos, sino que también aproxima las inercias de los yacimientos y los tipos. En el mapa de los yacimientos, por ejemplo, la inercia es aproximada (por abajo, como de costumbre) por la distancia del yacimiento al origen del mapa. O, equivalentemente, por la longitud del vector que corresponde con el yacimiento. Esto significa que un yacimiento que difiere muy poco de un yacimiento medio, y por lo tanto tiene una pequeña inercia, estará cerca del origen del mapa. Y los yacimientos que son diferentes de los otros tienden a estar en la periferia del mapa. Como consecuencia de ello puede ocurrir muy fácilmente que el centro del mapa, la zona próxima al origen, tenga una gran aglomeración de yacimientos que sean similares al yacimiento medio.

Un programa de CA (usamos De Leeuw y Mair [2008b]) típicamente toma la matriz de abundancias y la dimensionalidad deseada del mapa como sus argumentos. Entonces produce como resultado coordenadas para los mapas de los objetos en las filas (yacimientos) y los objetos en las columnas (tipos). Además, puede proporcionar una variedad de gráficos, y proporciona una *descomposición de la inercia*. Este tipo de descomposición es muy familiar en PCA. Tomando la longitud al cuadrado ponderada de las proyecciones de los puntos de yacimientos en la primera dimensión, en la segunda dimensión, y así sucesivamente. Esto descompone la inercia total de los vectores en un componente debido a la primera dimensión, a la segunda dimensión, y así sucesivamente. Al dividir los componentes del total, podemos decir que un cierto porcentaje de la inercia se "explica" por la primera dimensión, otra más pequeña, porcentaje en la segunda dimensión, y así sucesivamente. En última instancia hay $t=\min(r-1,c-1)$ dimensiones, y cada uno de ellas se encarga de un porcentaje decreciente determinado de la inercia total.

El CA también puede hacer *mapas conjuntos* o biplots, en los que, básicamente, tomamos el mapa de yacimientos y el de tipos y los ponemos uno encima del otro. Entonces tenemos un mapa en el que los tipos tenderán a estar cerca de los yacimientos en los que se producen con más frecuencia que cabría esperar sobre la base de los marginales. Decimos "tienden a", porque no hay distancia Benzécri definida entre un yacimiento y un tipo, y por lo tanto no hay aproximación en cierto sentido matemático bien definido. El programa de CA básicamente permite al usuario elegir entre cuatro opciones para el mapa conjunto.

La primera opción es poner los dos gráficos de Benzécri uno sobre el otro. Las distancias entre los yacimientos, y las distancias entre los tipos, se aproximan a las

distancias de Benzecri, pero las distancias entre yacimientos y tipos no pueden ponerse en relación con los datos de un modo simple. La segunda opción, que es llamada escalamiento de Goodman en el programa, es ajustar la longitud del yacimiento y los vectores de los tipos de tal manera que su producto interno se aproxime al residual de Pearson. Lamentablemente esto invalida la interpretación de las distancias entre yacimientos y tipos en términos de aproximaciones a las distancias de Benzécri. Las dos últimas opciones usan el *principio del centroide*. Podemos tomar el mapa de Benzécri para los yacimientos y, a continuación hacer gráficas de los tipos mediante la adopción de medias ponderadas (centroides) de los sitios, utilizando las frecuencias de los tipos en los yacimientos como ponderaciones. Esto produce un gráfico conjunto en el que las distancias entre los yacimientos aproximan las distancias de Benzécri. Las ubicaciones de los tipos en el gráfico de nuevo sólo difieren en longitud del vector de las localizaciones en el gráfico de tipos Benzécri. Las distancias entre los tipos ya no pueden interpretarse como aproximando las distancias de Benzécri entre los tipos, pero tienen una interpretación geométrica clara como medias ponderadas de los puntos del yacimiento. Por simetría hay un segundo principio de centroide, en la que utilizamos el gráfico de tipo Benzécri y, a continuación representamos los yacimientos como medias ponderadas de los tipos.

El principio de centroide también se puede utilizar para ajustar yacimientos o tipos pasivos en los gráficos. Supongamos que un yacimiento adicional, no utilizado en el análisis, es excavado, y los objetos son clasificados utilizando la misma tipología que la utilizada en el análisis. Entonces, las puntuaciones de los tipos del análisis se pueden utilizar para calcular la puntuación para este yacimiento adicional nuevo, simplemente calculando la puntuación media de CA del yacimiento en cada una de las dimensiones. De la misma manera uno podría añadir tipos adicionales al análisis, utilizando las puntuaciones de yacimientos, por ejemplo si uno decide dividir un tipo original en dos tipos nuevos. Naturalmente, una alternativa es repetir la CA con los yacimientos adicionales y tipos, lo cual entonces permite determinar activamente la solución CA completa.

5.1 Kelley

Pasaremos a ilustrar el CA exploratorio con el ejemplo de Kelley, el de tamaño pequeño. Los mapas bidimensionales de los yacimientos y los tipos según el CA están en la figura 1.

Insertar la figura 1 por aquí

Como nos esperábamos, en el mapa de los yacimientos vemos tres grupos de puntos en los vértices de un triángulo. Como ya sabemos, el mapa unidimensional es simplemente la proyección de todos los puntos sobre el eje horizontal.

Insertar la figura 2 por aquí

En la Figura 2a vemos la aproximación a las distancias de Benzécri entre yacimientos en una dimensión, y en la figura 2b en dos dimensiones. Las distancias de Benzécri están en el eje horizontal, las distancias euclidianas en el eje vertical. Aproximación desde abajo significa que todos los puntos están por debajo de la línea de 45 grados de ajuste perfecto. Pero, como podemos ver, el ajuste en dos dimensiones es ya casi perfecto. En cambio, en sólo una dimensión algunas de las distancias Benzécri más

grandes, en particular aquellas entre (21,34) y (23,37) están muy infraestimadas.

Mostramos finalmente la descomposición de chi-cuadrado para el ejemplo de Kelley. Como se podía esperar, las dos primeras dimensiones representan el 97% de la inercia total, y la tercera dimensión es de muy poca importancia.

Insertar Tabla 9 por aquí

5.2 Kolomoki

Aplicaremos ahora el CA a los datos de Kolomoki, nuestro ejemplo más realista. La descomposición chi-cuadrado se da en la Tabla 10. Dos dimensiones explican el 80% de la inercia, tres dimensiones casi el 90%. Los mapas de CA para los tipos en dos y tres dimensiones se dan en la Figura 3 y en la Figura 4. Una vez más, el mapa de dos dimensiones es sólo la proyección del mapa tridimensional en el plano horizontal (a excepción de una posible rotación). Se debe tener en cuenta que los puntos en los mapas bidimensionales son el centro de elipses de tamaños variables. Estas elipses son regiones al 95% de confianza para los puntos. El cálculo de las regiones de confianza, el cual es hecho por De Leeuw y Mair [2008b], se basa en el supuesto de que las abundancias son una gran muestra aleatoria de una población. Al igual que con la chi-cuadrado, esta hipótesis puede no ser apropiada en los ejemplos arqueológicos, pero, también como con la chi-cuadrado, el tamaño de las elipses da una representación útil de la variabilidad. Vemos grandes elipses para los puntos de la periferia, los cuales generalmente se corresponden con abundancias menores, y vemos ejemplos de elipses superpuestas para los yacimientos o tipos que realmente no pueden ser diferenciados.

Insertar la figura 3 por aquí

Para la interpretación de los resultados de dos dimensiones de Kolomoki, nos referimos a los expertos Smith y Neiman [2007]. La tercera dimensión no añade mucho (sólo el 9% de la inercia total), pero permite una mejor aproximación a algunas de las distancias Benzécri más grandes. En particular, la tercera dimensión enfatiza las diferencias entre los valores extremos T9 y (T1, T18).

Insertar la Figura 4 por aquí

Si seguimos añadiendo dimensiones, probablemente veremos cómo cada nueva dimensión se ocupa de un grupo de distancias Benzécri grandes, las cuales están todavía muy infraestimadas en tres dimensiones.

Insertar la figura 5 por aquí

Insertar Tabla 10 por aquí

6 Preguntas frecuentes

Existen diferentes versiones del CA que surgen de modo natural. No las hemos aplicado en nuestro ejemplo, pero las mencionaremos brevemente por mor de ser comprehensivos. Así, uno puede preguntarse por ejemplo, si la aproximación desde abajo es realmente una buena idea. Parece obvio que una mejor aproximación a las

distancias de Benzécri es posible si permitimos que algunas de las distancias del mapa sean sobreestimadas, y otras infraestimadas. Esta idea es explotada en [De Leeuw y Meulman, 1986]. La idea es, básicamente, calcular distancias de Benzécri primero, y luego aplicar escalamiento multidimensional a estas distancias.

Una segunda pregunta sería si existen alternativas adecuadas a las distancias de Benzécri. Recordemos que las distancias de Benzécri se utilizan porque corregimos las proporciones por sus errores típicos, bajo el supuesto de independencia. Las distancias de Benzécri tienen una conexión natural con chi-cuadrado, la suma ponderada de cuadrados, y por lo tanto con la distancia euclídea. Métodos alternativos para ponderar las proporciones son sin duda posibles, como en el CA esférico de Domingues y Volle [1980], pero en general la conexión con la geometría euclidiana se hace menos transparente.

Y, finalmente, podemos alejarnos de la interpretación de las matrices de abundancia en términos de frecuencias relativas. En su lugar, podemos pensar en ellas como datos sobre *datos composicionales*. Cada fila es un vector de proporciones, que sumado da uno, pero las proporciones pueden provenir de un análisis químico de muestras, y no venir de recuentos. Los datos composicionales son muy comunes en Quimiometría y Ciencias de la Tierra, y también bastante comunes en Arqueología. Variaciones de análisis de componentes principales para datos composicionales, similares a pero no idénticos al CA, se analizan en la monografía de Aitchison [2003].

7 Modelos de Distancia Exponencial

En Ecología [Ihm y van Groenewoud, 1975, Ter Braak, 1985], y, en cierta medida en Arqueología, se ha prestado mucha atención al modelo de Ordenación Gaussiana (Gaussian Ordination Model-GOM). El modelo dice que para el yacimiento *i* y tipo *j* el valor esperado de la abundancia es

$$\mathbf{E}(f_{ij}) = \alpha_i \beta_j \exp(-\frac{1}{2}\left(\frac{x_i - y_j}{s_j}\right)^2).$$

Así, los yacimientos y los tipos pueden ser escalados sobre una escala unidimensional común. La abundancia f_{ij} está, excepto por los efectos marginales de fila y columna α_i y β_j, relacionada con la distancia entre el valor de escala del yacimiento *i* y el valor de escala del tipo *j*. Más precisamente, un tipo será abundante en los yacimientos cuyo valor de escala es cercano al valor de escala del tipo, y será el más grande si el tipo y el yacimiento coinciden sobre la escala. Las filas de la matriz de abundancia serán unimodales: tienen un solo pico y luego descienden en ambas lados. O, utilizando la terminología de Kendall, son Q-matrices. Una vez más, excepto por los efectos marginales, lo mismo es cierto para las columnas. Así, si el modelo ajusta, podemos reordenar los yacimientos y los tipos de tal manera que tanto las filas y las columnas de la matriz de abundancia sean unimodales.

El GOM puede ser generalizado fácilmente a más de una dimensión.

$$\mathbf{E}(f_{ij}) = \alpha_i \beta_j \exp(-\frac{1}{2}\sum_{s=1}^{p} (x_{is} - y_{js})^2).$$

Por razones obvias, llamamos a esto el Modelo de Distancia Exponencial (EDM). El EDM es unimodal en un sentido geométrico más general. Las curvas de respuesta en el plano, si p=2, tienen un solo pico y descienden en todas las direcciones. Hay muchas maneras en las que el EDM puede ser ajustado a matrices de abundancia. La mayoría de ellos se basan en máxima verosimilitud multinomial, y por lo tanto de forma natural están acompañados con pruebas de significación en grandes muestras y regiones de confianza. Como no es de extrañar, ha habido contribuciones tanto desde el punto de vista de la psicometría como de la ecología. Para una técnica propuesta recientemente, y una buena revisión de trabajos previos, remitimos a De Rooij y Heiser [2005].

Podemos simplificar el EDM, expandiendo el cuadrado y juntando términos, en la forma equivalente

$$\mathbf{E}(f_{ij}) = \alpha_i \beta_j \exp(\sum_{s=1}^{p} x_{is} y_{js}).$$

Esto muestra que expandimos las abundancias en el producto de los efectos marginales y un término de interacción, el cual es el producto interno de los efectos de fila y columna. Esto está realmente muy cerca del CA. Para argumentos pequeños tenemos $\exp(x) \approx 1+x$, y, por tanto

$$\mathbf{E}(f_{ij}) \approx \alpha_i \beta_j (1 + \sum_{s=1}^{p} x_{is} y_{js}).$$

Este es el modelo que es ajustado por CA, utilizando mínimos cuadrados ponderados. Así, vemos que CA puede interpretarse como una aproximación conveniente y económica de la EDM, pero también como un modelo con derecho propio en el que las interacciones multiplicativas (exponencial) son sustituidas por unas aditivas. Además de esto, por supuesto, tanto EDM y CA pueden ser discutidos como métodos de reducción de datos y métodos de representación de datos, sin hacer necesariamente referencia a un modelo estadístico.

Las dos dimensiones Kolomoki de la solución de EDM se muestran en la figura 6. No daremos una interpretación del resultado, sino simplemente señalaremos que existen algunas diferencias con la solución de CA. La agrupación de los yacimientos y los tipos es aproximadamente la misma, pero la solución EDM muestra menos herradura, y eso es habitual.

Insertar la Figura 6 por aquí

8 Discusión

Este capítulo se podría llamar "las muchas caras del Análisis de Correspondencias". Intenta proporcionar diversos marcos de interpretación para mirar a los gráficos del CA, en términos de distancias, centroides, modelos de asociación, y chi cuadrado. También muestra cómo los mismos modelos y técnicas aparecen en muchas disciplinas diferentes, a menudo bajo diferentes nombres, y que combinar ideas de estas disciplinas

da posibilidades adicionales de interpretación.

También hemos discutido el modelo EDM, en sus diversos disfraces tal y como el GOM o el modelo RC. Puede ser utilizado para incorporar una forma de CA dentro un marco de máxima verosimilitud y cambiar el énfasis en la exploración multivariada a la comprobación de modelos.

Los arqueólogos no familiarizados con el CA pueden utilizar este capítulo para ver ejemplos anteriores en su disciplina, y pensar de manera diferente acerca de las matrices de abundancia y de incidencia. Hemos tratado de poner de relieve la continuidad entre el CA y los métodos de seriación anterior utilizados en arqueología.

Como hemos indicado, existen paquetes gratuitos en `R` disponibles para CA. Hemos mencionado `homals` y `Anacor`, pero en De Leeuw y Mair [2008b] se discuten también otros paquetes disponibles. Todos los sistemas estadísticos, tal y como `SAS`, `SPSS`, `Stata`, tienen métodos de CA bien como parte de ellos mismos o como módulos añadidos.

Referencias

J. Aitchison. *The Statistical Analysis of Compositional Data*. The Blackburn Press, Caldwell, New Jersey, 2003.

B.V. Arlif. The Archeological Seriation Problem. Master's thesis, Instutute of Computer Science, University of Copenhagen, Denmark, 1995.

M.J. Baxter. *Exploratory Multivariate Analysis in Archeology*. Edinburgh University Press, Edinburgh, 1994.

E. Beh. Simple Correspondence Analysis: A Bibliographic Review. *International Statistical Review*, 72:257–284, 2004.

J.P. Benzécri. *Analyse des Données: Taxonomie*, volume 1. Dunod, Paris, 1973a.

J.P. Benzécri. *Analyse des Données: Correspondances*, volume 2. Dunod, Paris, 1973b.

E. Bølviken, E. Helskog, K. Helskog, I.M. Holm-Olsen, L. Solheim, and R. Bertelsen. Correspondence Analysis: an Alternative to Principal Components. *World Archeology*, 14:41–60, 1982.

C.C. Clogg and E.S. Shihadeh. *Statistical Models for Ordinal Variables*. Number 4 in Advanced Quantitative Techniques in the Social Sciences. Sage Publications, Thousand Oaks, CA, 1994.

R.A. Clouse. Interpreting Archeological Data Through Correspondence Analysis. *Historical Archeology*, 33:90–107, 1999.

C. H. Coombs. *A Theory of Data*. Wiley, 1964.

J. De Leeuw. Nonlinear Principal Component Analysis and Related Techniques. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*, pages 107-133. Chapman and Hall, 2006.

J. De Leeuw. On the Prehistory of Correspondence Analysis. *Statistica Neerlandica*, 37:161–164, 1983.

J. De Leeuw and P. Mair. Homogeneity Analysis in R: The package homals. *Journal of Statistical Software*, 2008a.

J. De Leeuw and P. Mair. Simple and Canonical Correspondence Analysis Using the R Package anacor. *Journal of Statistical Software*, 2008b.

J. De Leeuw and J.J. Meulman. Principal Component Analysis and Restricted Multidimensional Scaling. In W. Gaul and M. Schader, editors, *Classification as a*

Tool of Research, pages 83–96, Amsterdam, London, New York, Tokyo, 1986. North-Holland.

M De Rooij and W.J. Heiser. Graphical Representations and Odds Ratios in a Distance Association Model for the Analysis of Cross-Classified Data. *Psychometrika*, 70:99–122, 2005.

D. Domingues and M. Volle. L'Analyse Factorielle Sphérique. In E. Diday, L. Lebart, J. Pagès, and R. Tomassone, editors, *Data Analysis and Informatics*, volume I, Amsterdam, Netherlands, 1980. North Holland Publishing Company.

A.I. Duff. Ceramic Micro-Seriation: Types or Attributes. *American Antiquity*, 61:89–101, 1996.

J.A. Ford. Measurements of Some Prehistoric Design Developments in the Southeastern States. *Anthropological Papers of the American Museum of Natural History*, 44(3):313–384, 1952.

H.G. Gauch, Jr. *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge, U.K., 1982.

A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, England, 1990.

Z. Gilula and S.J. Haberman. Canonical Analysis of Contingency Tables by Maximum Likelihood. *Journal of the American Statistical Association*, 81: 780–788, 1986.

L.A. Goodman. Simple Models for the Analysis of Association in Cross-classifications Having Ordered Categories. *Journal of American Statistical Association*, 74: 537–552, 1979.

J.C. Gower and D.J. Hand. *Biplots*. Number 54 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1996.

M. Greenacre and J. Blasius, editors. *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, 2006.

L. Guttman. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, editor, *The Prediction of Personal Adjustment*, pages 321–348. Social Science Research Council, New York, 1941.

L. Guttman. A Basis for Scaling Qualitative Data. *American Sociological Review*, 9:139–150, 1944.

L. Guttman. The Principal Components of Scale Analysis. In S.A. Stouffer and Others, editors, *Measurement and Prediction*. Princeton University Press, Princeton, 1950.

M.O. Hill. Correspondence Analysis: a Neglected Multivariate Method. *Applied Statistics*, 23:340–354, 1974.

F.R. Hodson, D.G. Kendall, and P. Tăutu, editors. *Mathematics in the Archeological and Historical Sciences*, Edinburgh, 1971. Edinburgh University Press.

P. Ihm and H. van Groenewoud. A Multivariate Ordering of Vegetation Data Based on Gaussian Type Gradient Response Curves. *The Journal of Ecology*, 63:767–777, 1975.

I. Kelley. *The Archeology of the Autlán-Tuxcacuesco Area of Jalisco. I: The Autlán Zone*, volume 26 of *Ibero-Americana*. University of California Press, 1945.

D.G. Kendall. Incidence Matrices, Interval Graphs, and Seriation in Archeology. *Pacific Journal of Mathematics*, 28:565–570, 1969.

D.G. Kendall. Abundance Matrices and Seriation in Archeology. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 17:104–112, 1971.

J.B. Kruskal. Multi-dimensional Scaling in Archeology: Time is not the Only Dimension. In F.R. Hodson, D.G. Kendall, and P. Tăutu, editors, *Mathematics in the Archeological and Historical Sciences*, pages 119–132, Edinburgh, 1971. Edinburgh University Press.

S.A. LeBlanc. Micro-Seriation: A Method for Fine Chronologic Differentiation. *American Antiquity*, 40:22–38, 1975.

J. Müller and A. Zimmerman, editors. *Archeology and Correspondence Analysis. Examples, Questions, Perspectives.*, volume IA 23 of *Internationale Archäologie*. Verlag Marie Leidorf, Rahden, Germany, 1997.

C. Orton. Plus ça Change ? 25 Years of Statistics in Archeology. In L. Dingwall, S. Exon, V. Gaffney, S. Laflan, and M. van Leusen, editors, *Archeology in the Age of the Internet*, Oxford, 1999. Archopress.

T.J. Pluckhahn. *Kolomoki: Settlement, Ceremony, and Status in the Deep South, A.D. 350-750*. University of Alabama Press, Tuscaloosa, Alabama, 2003.

J. Poblome and P.J.F. Groenen. Constrained Correspondence Analysis for Seriation of Sagalassos Tablewares. In M. Doerr and A. Sarris, editors, *Computer Applications and Quantitative Methods in Archaeology*, pages 301–306. Hellinic Ministry of Culture, 2003.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL `http://www.R-project.org`.

W.S. Robinson. A Method for Chronologically Ordering Archeological Deposits. *American Antiquity*, 16:293–301, 1951.

B.F. Schriever. *Order Dependence*. PhD thesis, University of Amsterdam, The Netherlands, 1985. Also published in 1985 by CWI, Amsterdam, The Netherlands.

B.F. Schriever. Scaling of Order-dependent Categorical Variables with Correspondence Analysis. *International Statistical Review*, 51:225–238, 1983.

W.H. Sears. *Excavations at Kolomoki: Final Report*. University of Georgia Press, Athens, Georgia, 1956.

K.Y. Smith and F.D. Neiman. Frequency Seriation, Correspondence Analysis, and Woodland Periodic Ceramic Assemblage Variation in the Deep South. *Southeastern Archeology*, 26(1):47–72, 2007.

C.J.F. Ter Braak. Correspondence Analysis of Incidence and Abundance Data: Porperties in Terms of a Unimodal Response Model. *Biometrics*, 41:859–873, 1985.

J.L.A. Van Rijckevorsel. *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. PhD thesis, University of Leiden, The Netherlands, 1987. Also published in 1987 by DSWO-Press, Leiden, The Netherlands.

R.H. Whittaker, editor. *Ordination of Plant Communities*. Dr. W. Junk BV, The Hague, Netherlands, 1978.

T.W. Yee. A New Technique for Maximum Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, 74:685–701, 2004.

Tabla 1: Matriz de Abundancias de Kelley

	Tipo				
	AutPol	MiReBr	AuWhRe	AltRed	
21	8	14	0	0	22
34	19	35	0	0	54
23	138	6	0	1	145
37	299	11	0	2	312
9	102	12	22	271	407
7	34	14	59	246	520
	600	92	81	520	1293

Tabla 2: Matriz de Proporciones de Kelley

	Tipo				
	AutPol	MiReBr	AuWhRe	AltRed	
21	0.006	0.011	0.000	0.000	0.017
34	0.015	0.027	0.000	0.000	0.041
23	0.107	0.005	0.000	0.001	0.112
37	0.231	0.009	0.000	0.002	0.241
9	0.079	0.009	0.017	0.210	0.315
7	0.026	0.011	0.046	0.190	0.273
	0.464	0.071	0.063	0.402	1.000

Tabla 3: Residuales de Pearson Kelley

	Tipo			
	AutPol	MiReBr	AuWhRe	AltRed
21	-0.02	+0.28	-0.03	-0.08
34	-0.03	+0.44	-0.05	-0.13
23	+0.24	-0.04	-0.08	-0.21
37	+0.35	-0.07	-0.12	-0.31
9	-0.18	-0.09	-0.02	+0.23
7	-0.28	-0.06	+0.22	+0.24

Tabla 4: puntuaciones z de Kelley

	Tipo			
	AutPol	MiReBr	AuWhRe	AltRed
21	-0.69	+9.94	-1.17	-2.97
34	-1.21	+15.90	-1.83	-4.66
23	+8.62	-1.34	-3.01	-7.50
37	+12.81	-2.38	-4.42	-11.02
9	-6.32	-3.15	-0.69	+8.39
7	-10.14	-2.22	+7.84	+8.73

Tabla 5: Condicionando sobre las filas en Kelley

	Tipo				
yacimiento	AutPol	MiReBr	AuWhRe	AltRed	$X^2_{i\bullet}$
21	0.36	0.64	0.00	0.00	4.98
34	0.35	0.65	0.00	0.00	0.04
23	0.95	0.04	0.00	0.01	0.11
37	0.96	0.03	0.00	0.01	0.24
9	0.25	0.03	0.05	0.52	0.31
7	0.10	0.04	0.17	0.47	0.27
$p_{\bullet j}$	0.46	0.07	0.06	0.40	0.93

Tabla 6: Condicionando sobre las columnas en Kelley

	Tipo				
yacimiento	AutPol	MiReBr	AuWhRe	AltRed	$p_{i\bullet}$
21	0.01	0.15	0.00	0.00	0.02
34	0.03	0.38	0.00	0.00	0.04
23	0.23	0.07	0.00	0.00	0.11
37	0.50	0.12	0.00	0.00	0.24
9	0.17	0.13	0.27	0.52	0.31
7	0.06	0.15	0.73	0.47	0.27
$X^2_{\bullet j}$	0.64	4.06	1.18	0.68	0.93

Tabla 7: Distancias de Benzécri al cuadrado para las filas (yacimientos)

	21	34	23	37	9	7
21	0.000					
34	0.002	0.000				
23	5.721	5.950	0.000			
37	5.841	6.072	0.001	0.000		
9	6.353	6.550	2.188	2.208	0.000	
7	6.812	6.999	3.207	3.233	0.259	0.000

Tabla 8: Distancias de Benzécri al cuadrado entre columnas (Tipos)

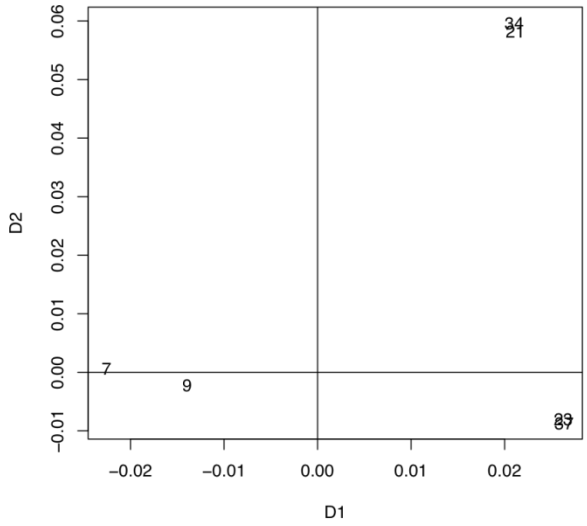
	AutPol	MiReBr	AuWhRe	AltRed
AutPol	0.000			
MiReBr	4.921	0.000		
AuWhRe	3.221	6.203	0.000	
AltRed	2.539	5.780	0.436	0.000

Tabla 9: Descomposición Chi-cuadrado para Kelley

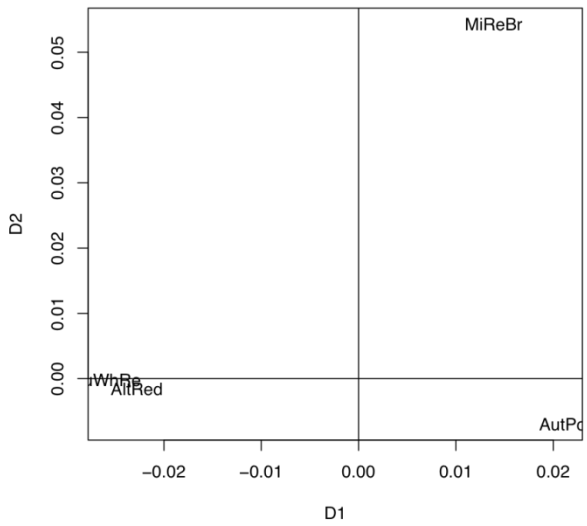
	X^2	%	Cum %
1	787.9	0.65	0.65
2	390.0	0.32	0.97
3	29.6	0.03	1.00
Total	1207.5		

Tabla 10: Descomposición Chi-cuadrado para Kolomoki

	X^2	%	Cum %
1	1018.8	0.63	0.63
2	261.6	0.16	0.79
3	144.7	0.09	0.88
4	128.0	0.08	0.96
5	38.6	0.02	0.98
6	17.9	0.01	0.99
7	9.0	0.01	1.00
8	3.8	0.00	1.00
Total	1622.5		



(a) Yacimientos

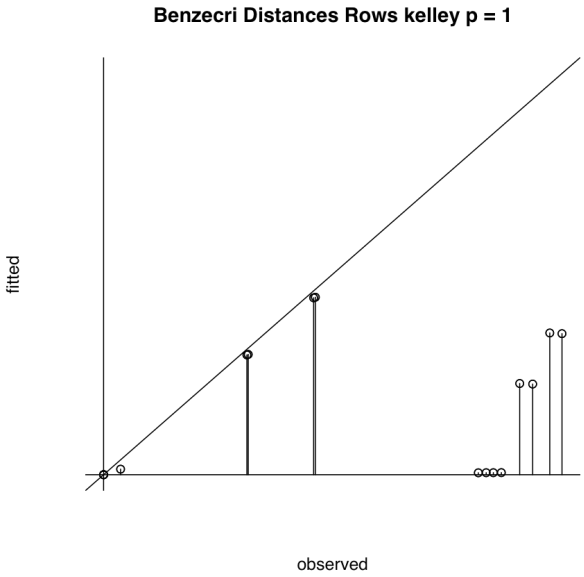


(b) Tipos

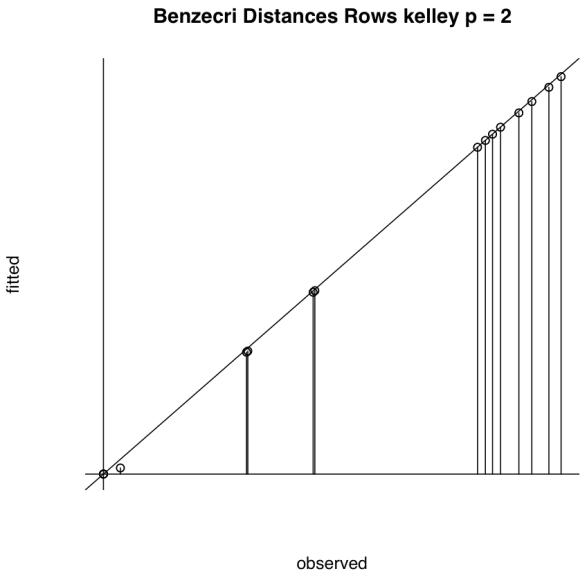
Figura 1: Mapa bidimensional de CA para Kelley

(a) Una Dimensión

Distancias de Benzecri entre las filas Kelley p =1
Ajustadas
Observadas



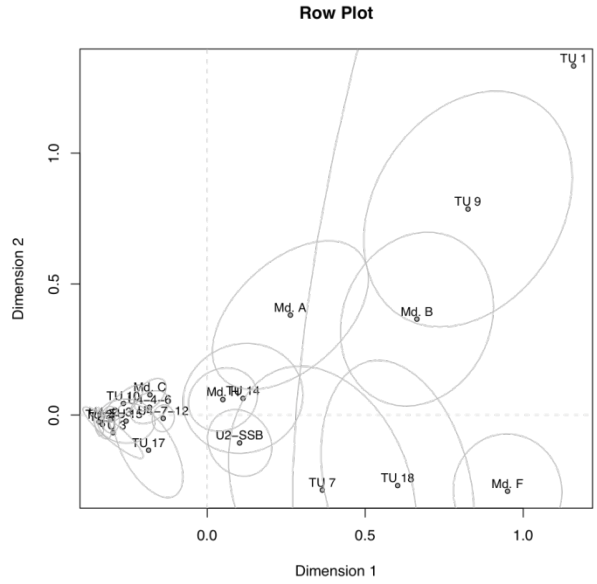
Distancias de Benzecri entre las filas keley p=2
Ajustadas
Observadas



(b) Dos Dimensiones

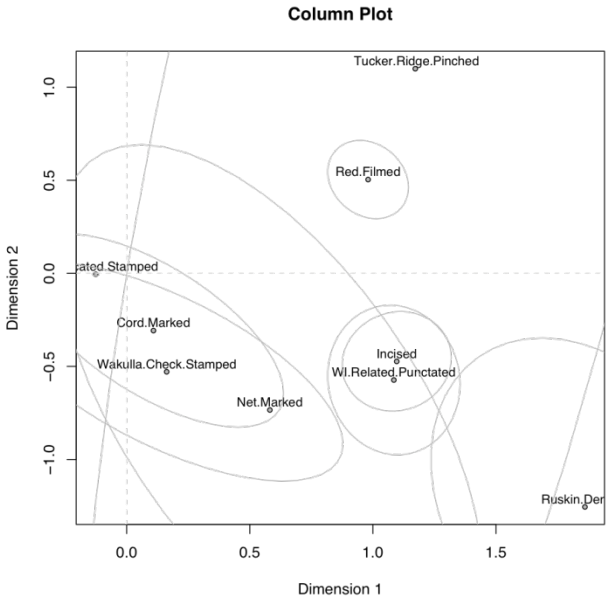
Figura 2: Aproximación a las Distancias de Benzécri para Kelley

Gráfico de filas



(a) filas

Gráfico de columnas



(b) Columnas

Figura 3: Mapas de CA de Kolomoki

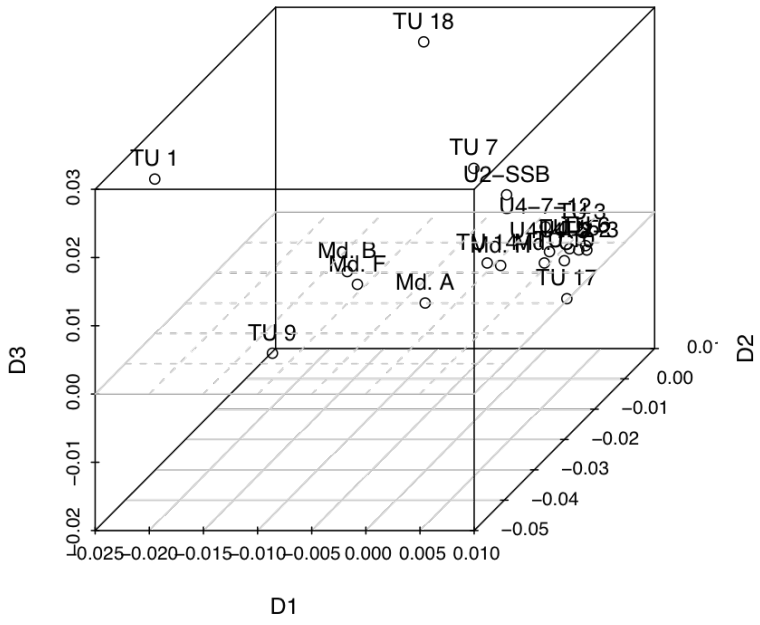
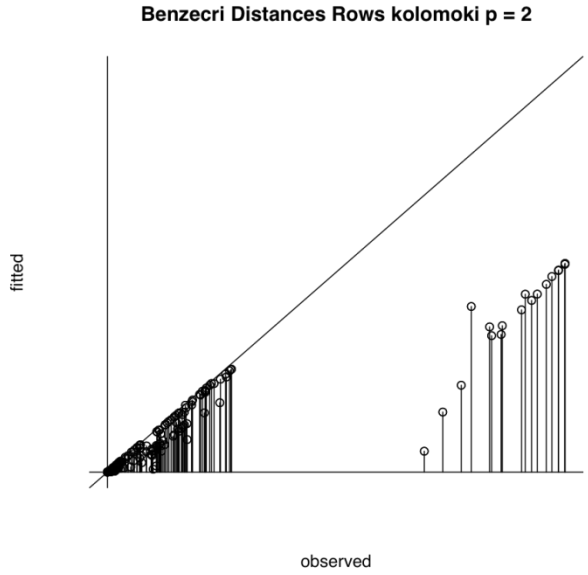


Figura 4: Mapa en tres dimensiones de Kolomoki

Distancias de Benzécri entre las filas kolomoli p=2

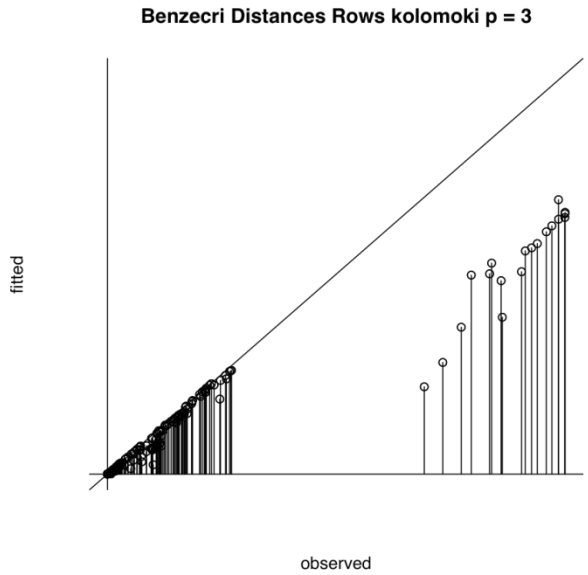
Ajustadas

Observadas



(a) Dos Dimensiones

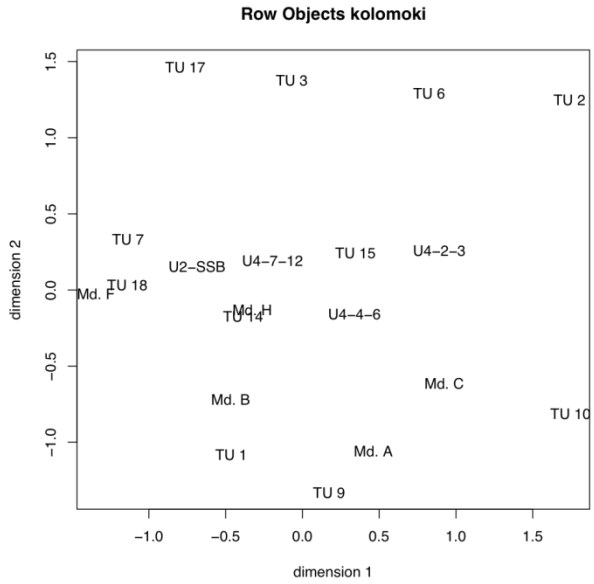
Distancias de Benzecri entre filas kolomoki p=3



(b) Tres Dimensiones

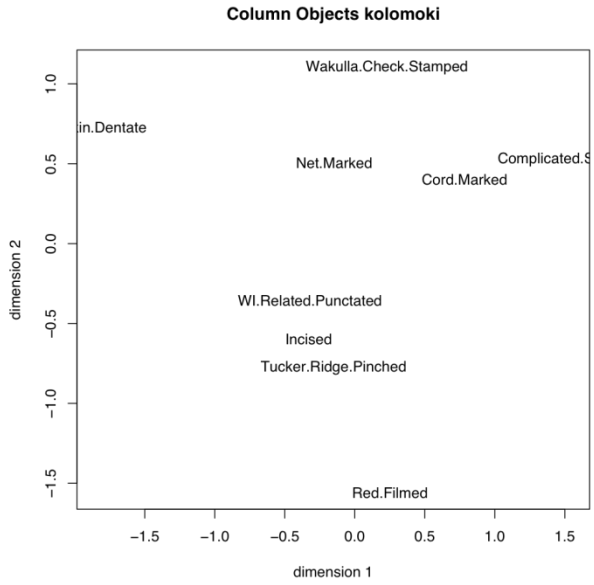
Figura 5: Aproximación a las distancias de Benzecri para Kolomoki

Objetos fila kolomoki



(a) Filas

Objetos columna kolomoki



(b) Columns

Figura 6: Mapas EDM de Kolomoki