EL USO DE UN METODO NUMERICO PARA LA DETERMINACION DE REGIONES BIOGEOGRAFICAS

Exequiel Ezcurra

Fundación Bariloche
Departamento de Recursos Naturales y Energía
C.C. 138 - 8400 San Carlos de Bariloche

RESUMEN

Los datos binarios (presencia-ausencia) obtenidos de los mapas de distribución geográfica de insectos-plaga de la agricultura se procesaron mediante un método divisivomonotético de clasificación numérica basado en el Estadístico de Información, con el fin de determinar regiones geográficas homogéneas. El método es particularmente adecuado para el procesamiento de grandes matrices de información binaria, por su simplicidad operativa y el poco tiempo de computación que requiere. La aproximación del Estadístico de Información a la función de χ^2 permite probar estadísticamente la significación de la clasificación resultante.

Se presenta una descripción detallada del método y se presentan los resultados en forma de un mapa del mundo dividido en 10 regiones, con el dendrograma correspondiente y un listado de las especies indicadoras utilizadas para cada división.

Se discute detalladamente las características del método y se analizan sus posibilidades de uso en la clasificación de información Biogeográfica y Ecológica.

SUMMARY

Binary data (presence-absence) taken from geographical distribution maps of agricultural insect-pests were processed with a divisive-monothetic method of numerical taxonomy, based on the Information Statistic (I), with the object of limiting homogeneous biogeographical regions. The method is particulary fit for the processing of large matrices of binary data, it is operationally simple and requires little computer time. The χ^2 approximation of the Information Statistic permits testing the Statistical significance or the resulting hierarchy.

A detailed description of the method is presented, as well as the results in the forma of a world map divided in 10 regions, with the corresponding dendrogram and a list of the discriminant species used in each division.

The characteristics of the method are discussed, and the possibilities of using it in the classification of Biogeographical and Ecological data are analysed.

Keywords: Numerical Taxonomy, Biogeography, Insect Pests.

I. INTRODUCCION

Varios autores han utilizado técnicas de clasificación numérica en problemas biogeográficos (Jardine, 1972; Kikkawa, 1968; Kikkawa y Pearde, 1969; Rapoport, Ezcurra y Drausal, 1976), como métodos relativamente rápidos que permiten clasificar áreas en forma objetiva, utilizando grandes cantidades de información.

Los datos que se procesan en estos problemas son de tipo cualitativo (presencia-ausencia), ya que se obtienen de los mapas de áreas de distribución. El problema no es exclusivo de la Biogeografía; es también común en otros campos de la Ecología y en el estudio de Recursos Naturales contar con una gran masa de información cualitativa, donde sólo se mide la presencia o ausencia de determinados atributos sobre un grupo de unidades taxonómicas, sin importar la cantidad en que cada atributo se encuentra presente.

Mas que la complejidad de los cálculos, el factor limitante en este tipo de problema es la cantidad de información a procesar, generalmente muy grande. Los algoritmos más comunes de

clasificación numérica (Grigal y Goldstein, 1972) son en estos casos excesivamente lentos sino imposibles de utilizar por el costo en tiempo de máquina. El presente trabajo explica la aplicación de un método particularmente eficiente para la clasificación de grandes matrices de información binaria. Kikkawa y Pearse (1969) utilizaron esta metodología para analizar la distribución de 464 especies de aves terrestres observadas en 121 sitios del continente Australiano. Los grupos obtenidos mostraron una notable semejanza con las áreas biogeográficas propuestas anteriormente por otros autores a través de métodos clásicos. Esto hace particularmente interesante el uso del método en el análisis de problemas biogeográficos y de taxonomía numérica en general.

Los datos utilizados en este trabajo se tomaron de los mapas de áreas de distribución de 281 insectos-plaga de la agricultura, editados por el Commonwalth Institute of Entomology (C.I.E.), Estos insectos representan uno de los grupos biológicos más cosmopolitas. Su distribución, íntimamente ligada a la acción humana, es de particular interés en Biogeografía. Dado que su evolución y/o su forma de dispersión se hallan fuertemente influídas por factores antrópicos, es lógico pensar que se agrupan en regiones distintas a las que ocupan los organismos naturales (Rapport, Ezcurra y Drausal, 1976).

2.- METODO

Codificación de los datos. Para cuantificar el área de distribución de cada especie se superpuso sobre el mapa correspondiente una grilla de 10368 cuadrados contiguos (72 x 144 divisiones). Así cada especie está definida por un vector de 10368 elementos con información cualitativa, y hay tantos vectores como especies consideradas (281 en este caso). La información se pudo codificar numéricamente en forma de una matriz de 10368 cuadrados (o unidades taxonómicas) y 281 especies (o atributos), donde cada elemento (i, j) de la matriz vale cero si la especie j está ausente en el cuadrado i, y vale 1 si está presente.

Análisis de Asociación. Los métodos jerarquicos de clasificación numérica se dividen convencionalmente en métodos **aglomerativos** que unen unidades taxonómicas similares en conjuntos cada vez mayores, y los **divisivos** que progresivamente subdividen el conjunto de unidades taxonómicas en subconjuntos homogéneos. Los métodos aglomerativos son excesivamente lentos cuando la matriz de datos es grande, y presentan el inconveniente adicional de que pequeños errores en los datos pueden reunir unidades taxonómicas en grupos al que no corresponden fenómeno conocido como "efecto de encadenamiento" (Lance y Williams, 1965; Wishart, 1969).

En los casos en que los atributos se encuentran asociados en alguna medida, la clasificación divisiva monotética es aceptable y a veces la más apropiada, porque provoca menos errores (Lance y Williams, 1965). Este método de clasificación se conoce como "Análisis de Asociación" (Williams y Lambert, 1959) y es apto sólo y particularmente en los casos en que se posea información de tipo binaria o cualitativa. Dada una población definida por atributos binarios, se busca aquel atributo que al dividir la población según su presencia y ausencia crea dos grupos de máxima homogeneidad. Se demuestra (Lance y Williams, 1965) que el atributo elegido es el más fuertemente asociado con los restantes, ya sea en forma positiva o negativa, y que la partición que provoca es por lo tanto muy cercana al óptimo.

El Estadístico de Información. Para realizar el Análisis de Asociación es necesario medir la heterogeneidad del grupo de unidades taxonómicas. En este trabajo se emplea el Estadístico de Información (I) (Lance y Williams, 1968; Pielou, 1969).

Supongamos que el grupo consiste en **n** unidades taxonómicas (en nuestro caso especies), donde a_j es la cantidad de unidades taxonómicas que contienen el atributo j. Entonces, la proporción de unidades taxonómicas que contienen el atributo j es a_j/n, y la proporción de unidades que no lo contienen es $(n-a_j)/n$. Considerando que el grupo está formado por unidades de dos clases (las que contienen el atributo j y las que no lo contienen) podemos usar la fórmula de Shannon para determinar la información por unidad taxonómica (o "diversidad") respecto del atributo j.

$$H'_j = -\frac{a_j}{n}\log\frac{a_j}{n} - \frac{(n-a_j)}{n}\log\frac{(n-a_j)}{n}$$

La información (o heterogeneidad) **total** del grupo respecto del atributo j es entonces $n\,H'_j$. Sumando para todos los atributos se obtiene el Estadístico de Información (I), que representa la heterogeneidad del grupo respecto del total de atributos.

$$I = \sum_{j=1}^{s} n\,H'_j = -\sum_{j=1}^{S}\left[a_j \log\frac{a_j}{n} + (n-a_j)\log\frac{(n-a_j)}{n}\right]$$

$$I = sn\log n - \sum_{j=1}\left[a_j \log a_j + (n-a_j)\log(n-a_j)\right]$$

Las unidades del Estadístico de Información dependen de la base logarítmica utilizada. Si se utiliza $\log_2$, la unidad de Información es el bit; si se utiliza $\log_e$, la Información se mide en nats; y si se emplea $\log_{10}$ la medida es el Hartley (Abramson, 1963). En este trabajo, como en la mayoría de los casos, se utiliza la base e, pero es fácil la conversión de una unidad a otra multiplicando por el factor de conversión de sus respectivos logaritmos (e.g. 1 nat =1.44269 bits).

El programa de Clasificación

Si un conjunto de unidades taxonómicas (i) se divide en dos subconjuntos (g) y (h), la caída de Información $\Delta I_{(gh,i)}$, se define como:

$$\Delta I_{(gh,i)} = I_i - (I_g + I_h)$$

Para partir el conjunto, el programa divide el total de unidades taxonómicas de acuerdo a la presencia y ausencia de cada atributo sucesivamente, y elige el atributo que provoca un máximo Δ I como atributo discriminante. Una vez formados dos o más grupos elige el más heterogéneo (el de mayor valor de I) para la próxima división, y así sucesivamente hasta lograr el número de grupos deseado. Los valores de heterogeneidad (I) de cada área corresponden a su nivel jerarquico dentro del dendrograma.

La significación estadística de cada división se puede docimar tomando a $2\Delta I$ como un estimador sesgado de χ^2, con tantos grados de libertad como atributos discriminantes (Lance y Williams, 1968). Se considera atributo discriminante a aquel que se encuentra presente en algunas unidades del grupo y ausente en otros. Los atributos que están totalmente presentes o totalmente ausentes en el grupo son irrelevantes ya que no agregan heterogeneidad al conjunto.

El Programa de Clasificación (MAMOTRETO) fué escrito en FORTRAN con cinco subrutinas centrales en B.A.L. Los cuadrados sin especies (mares, desiertos y selvas) no entraron en la clasificación. Los resultados se obtuvieron en forma de mapa dibujado con la impresora de líneas donde cada dígito corresponde a un cuadrado, y áreas iguales aparecen con la misma numeración. El trabajo se procesó en una Honeywell-Bull 66.

3. RESULTADOS

Para cada división los resultados que arroja el computador son: los valores de heterogeneidad (I) de cada área, la especie elegida como discriminante óptima en el área más heterogénea, la cantidad de especies discriminantes en dicha área (equivalente a grados de libertad para la dócima de χ^2) y el mapa del mundo impreso en forma digital, con las distintas areas señaladas por dígitos diferentes. Usando la aproximación de χ^2, todas las particiones realizadas fueron altamente significativas ($P \leq 0{,}001$)

El tiempo empleado por el computador fue de aproximadamente 3 horas. Pruebas efectuadas con algoritmos aglomerativos de clasificación numérica nos permite estimar que el tiempo necesario hubiera sido mayor en uno o varios órdenes de magnitud, lo que hubiera hecho impracticable la resolución del problema.

En la figura 1 se presenta el mapa resultante hasta 10 divisiones y en la figura 2 el dendrograma correspondiente (tomadas de Ezcurra, Rapoport y Marino, 1977). Siguiendo las ramificaciones del dendrograma se puede observar como se produjeron las particiones anteriores. Los números en cada ramificación corresponden al número de identificación del C.I.E. para la especie que determina esa división. Dichas especies son: -9.- **Cydia pomonella (L) (=Carpocapsa pomonella L.)** (Lep., Tortricidae); -27.- **Nezara viridula** (L.) (Hemip., Pentatomidae); -39.- **Lygus prantesis (L.)** (Hemip., Miridae); -277.- **Xyleborus ferrugineus (F.)** (=X. Confusus Eich.) (Col., Scolytidae); -237.- **Sesamia inferens (Wlk).** (Lep. Noctuidae); -20.- **Thrips tabaci** (Lind.) (Tisanop. Thripidae); -38.- **Lygus oblineatus** (Say) (Hemip. Miridae); -239.- **Heliothis zea** (Boddie) (Lep., Noctuidade); y -69.- **Operophtera brumata** (L) (Lep., Geometridae).

Dado que el objetivo de este trabajo es explicar la aplicación de la metodología, no se desarrolla aquí un análisis Biogeográfico de los resultados, que hemos realizado en otro trabajo (Ezcurra, Rapoport y Marino, 1977). Es importante destacar sin embargo, como las áreas de características más similares se agrupan en el dendrograma. Las regiones 1, 2 y 3 forman un grupo homogéneo de regiones de escasa o ninguna agricultura. Las regiones 4, 5 y 6 abarcan zonas de agricultura tropical. Las regiones 7 y 8 abarcan zonas de agricultura templada de colonización reciente o relativamente extensivas (América, Sud Africa, Australia y Nueva Zelandia, y partes de Asia); mientras que las regiones 9 y 10 abarcan toda Europa y corresponden a zonas de agricultura templada tradicional e intensiva.

4. DISCUSION

Dado que el método de clasificación realiza las particiones de acuerdo con la presencia y ausencia de sólo un atributo, es fundamental que el atributo según el cual se realiza la partición esté asociado a los restantes, en forma po-

FIGURA 1:
Regiones Mundiales de distribución de insectos-plagas de la agricultura.

sitiva o negativa. Dicho de otra forma, para que las particiones sean casi óptimas es necesario que exista una marcada tendencia hacia la existencia de grupos de unidades taxonómicas con los mismos atributos representativos como indicadores de una división. Este argumento se ha usado frecuentemente como crítica a los métodos monotéticos, pero es importante destacar en defensa del método que cuando la distribución de los atributos no es agrupada, es un error pretender clasificar o agrupar las unidades por ningún método; lo que corresponde es utilizar metodologías de ordenación.

Como ventaja del método podemos mencionar que la aproximación de χ^2 permite docimar la bondad estadística de cada partición, y si $2 \Delta I$ da valores exceisvamente bajos para los Grados de Libertad se puede inferir que la matriz de datos posee una distribución no apropiada para clasificar, ya sea porque es suficientemente homogénea como para ser considerada un solo grupo "per se" (bajo valor de I), o porque la distribución de los datos es continua o al azar.

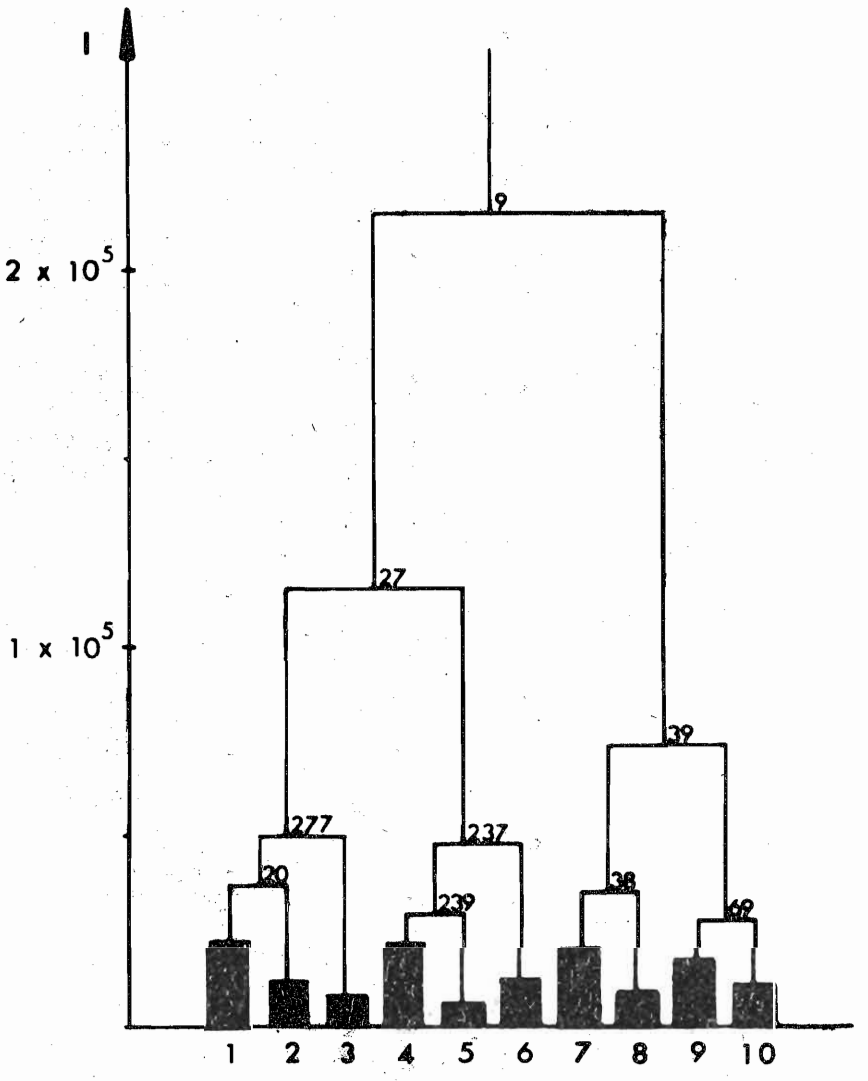

FIGURA 2:
Dendrograma. Los números en cada ramificación corresponden al número de identificación de la especie característica. La altura de las columnas en negro corresponde a la heterogeneidad de cada área.

Bottomley (1971), señala como inconveniente del método el que, dada su simetría de cálculo para presencia y ausencia, reúne en un mismo grupo a unidades de muy pocos atributos presentes, aunque los atributos no sean los mismos. Estas unidades se asemejan en que prácticamente valen cero para todos los atributos, más que en el hecho de compartir los pocos atributos presentes. En los resultados esto se nota en la región 1, que corresponde a zonas frías y desiertos. Los cuadrados que forman esta región se parecen en que casi no poseen insectos-plagas de la agricultura, pero no necesariamente comparten las pocas especies que se hallan en uno u otro lugar. En los casos de procesamientos en que se desee evitar este inconveniente, se deben procesar en cada grupo solo los atributos que tienen por lo menos una presencia dentro del grupo (Bottomley, 1971).

5. CONCLUSIONES

El uso del Análisis de Asociación es de gran utilidad en la determinación de regiones geográficas homogéneas a partir de áreas de distribución. Más que en la precisión de los resultados, la ventaja del método radica en que permite procesar grandes volúmenes de información que de otra manera no se podrían asimilar, y detectar, a grandes rasgos, la existencia de agrupamientos en los datos, con un cierto nivel de confianza estadístico.

El método es apto no solo para el procesamiento de información biogeográfica, sino también para la clasificación de comunidades en Ecología, la clasificación de datos o inventarios de Recursos Naturales; o en general para el agrupamiento de grandes masas de información binaria.

AGRADECIMIENTOS

Al personal del Centro de Cómputos de Fundación Bariloche, en especial a Harald Solberg y Eduardo Rodríguez, sin cuya ayuda este trabajo hubiera sido imposible.

Especial reconocimiento al Dr. E. H. Rapoport por todas las ideas, críticas y sugerencias que alentaron este trabajo y por la lectura del manuscrito.

BIBLIOGRAFIA

ABRAMSON, N. 1969. Teoría de la Información y Codificación. 2ª Ed. Paraninfo, España.

BOTTOMLEY, J. 1971. Some statistical problems arising from the use of the information statistic in numerical classification. J. Ecol. 59:339-342.

EZCURRA, E., RAPOPORT, E.H. y MARINO, C. The geographical distribution of insect pests. En prensa.

GRIGAL, D.F. y GOLDSTEIN, R.A. 1972. Computer programs for the ordination and classification of ecosystems. Publs. Ecol. Sci. Div. 147, 125.

JARDINE, N. 1972. Computational methods in the study of plant distributions, en Valentine, D.H. (Ed.) Taxonomy, Phytogeography and Evolution. Academic Press, Nueva York y Londres pp. 381-393.

KIKKAWA, J. 1968. Ecological association of birds species and habitats in eastern Australia; similarity analysis. J. Anim. Ecol. 37:143-165.

KIKKAWA, J. y PEARSE, K. 1969. Geographical distribution of land birds in Australia - a numerical analysis"). Aust. J. Zool. 17:821-840.

LANCE, G.N. y WILLIAMS, W.T. 1965. Computer programs for monothetic classification ("Association analysis"). Computer J.8:246-249.

LANCE, G.N. y WILLIAMS, W.T. 1968. Note on a new information-statistic classificatory program. Computer J. 11:195.

PIELOU, E.C. 1969. An Introduction to Mathematical Ecology. J. Wiley & Sons, Nueva York y Londres. 286 pp.

RAPOPORT, E.H., EZCURRA, E. y DRAUSAL, B. 1976. The distribution of plant diseases: a look into the biogeography of the future. J. Biogeogr. 3:1-8.

WILLIAMS, W.T. y LAMBERT, J.M. 1959. Multivariate methods in plant ecology, I. J. Ecol. 47:83-101.

WISHART, D. 1969. Mode Analysis: A generalisation of nearest neighbour which reduces chaining effects. En Cole, A.J. (Ed.) Numerical Taxonomy, Proc. Colloq. In Num. Tax., Univ. St. Andrews, Sept. 1968. Academic Press, Nueva York y Londres. pp. 282-311.