 ISSN 1993 8012

La paradoja de Simpson en la exploración de yacimientos lateríticos cubanos

Rafael Arturo Trujillo-Codorníu
Santiago Bernal-Hernández
Mirelis Rasúa-López

Resumen
En este artículo se reportan varios ejemplos de la paradoja de Simpson en la exploración de yacimientos lateríticos del oriente cubano y se discuten los efectos y las condiciones para la reversión de Simpson.

Palabras clave
Paradoja de Simpson; yacimientos lateríticos.

Recibido: 12 abril 2012 / Aceptado: 26 abril 2012

 ISSN 1993 8012

Simpson's paradox in resource evaluation of Cuban lateritic deposits

Abstract
In this paper, we report several examples of Simpson's paradox in resource evaluation of Cuban lateritic deposits and discuss effects and boundary conditions of the Simpson's reversal.

Keywords
Simpson's paradox; lateritic deposits.

Received: 12 April 2011 / Accepted: 26 April 2012

 ISSN 1993 8012

1. INTRODUCCIÓN

La paradoja de Simpson (o efecto de Yule-Simpson) es una paradoja en la cual una correlación existente en diferentes grupos es revertida cuando los grupos son combinados. Este efecto está presente, sobre todo, en las ciencias sociales y en la estadística médica. Uno de los ejemplos clásicos, en la estadística médica, es el reportado por Charig *et al*. (1986) y por Julious & Mullee (1994) referido a la tasa de éxito de un tratamiento para los cálculos renales. La Tabla 1 muestra la tasa de éxito y el número de pacientes atendidos con cálculos renales clasificados en dos categorías: pequeños y grandes, donde el tratamiento A se refiere a los tratamientos quirúrgicos y el B es la nefrolitotomía percutánea.

Tabla 1. Comparación de las tasas de éxito en dos tratamientos para los cálculos renales

	Tratamiento A	**Tratamiento B**
Pequeños cálculos	**93** % (81/87)	87 % (234/270)
Grandes cálculos	**73** % (192/263)	69 % (55/80)
Ambos	78 % (273/350)	**83** % (289/350)

La conclusión paradójica es que el tratamiento A es más efectivo tanto para cálculos pequeños como para cálculos grandes, pero aparece como menos efectivo cuando se consideran ambos casos de conjunto.

Como puede apreciarse, a menos que se use juiciosamente, la paradoja de Simpson puede provocar dificultades en la inferencia estadística. La sola idea de su existencia, de que, por ejemplo, un tratamiento que beneficia a hombres y mujeres por separado puede ser dañino a la población en su conjunto, muestra que muchas conclusiones intuitivas que se hacen cotidianamente no están soportadas por el cálculo de las probabilidades.

Recientemente se han reportado múltiples ejemplos de aparición de la paradoja de Simpson en disímiles esferas. Terwilliger & Schield (2004) han encontrado cerca de 100 instancias de la paradoja de Simpson en las estadísticas de los resultados escolares de los Estados Unidos. Similares estudios se han realizado en los resultados escolares de Israel (Zuzovsky & Steinberg 2011) y como dato curioso Ma & Ma (2011) muestran ejemplos de la paradoja de Simpson en las estadísticas de la liga de baloncesto norteamericana.

 ISSN 1993 8012

Uno de los primeros reportes, hasta donde conocemos, de la ocurrencia de la paradoja de Simpson en la evaluación de recursos naturales aparece en el trabajo de Ma (2009). En este trabajo se ofrecen varios ejemplos (ver uno de ellos en la Tabla 2), tomados esencialmente de la geología del petróleo y se discute además la relación entre la paradoja de Simpson, el problema de la inferencia ecológica (que ocurre cuando se utilizan datos agregados para llegar a conclusiones a nivel individual) y el problema del cambio de soporte.

Tabla 2. Porosidad media de dos formaciones geológicas y dos facies en un depósito de carbonatos en América del Norte (Ma 2009)

Formaciones	Facies de playa	Facies de aguas someras
Formación A	**11,10**%	10,59%
Formación B	**7,89**%	7,54%
En conjunto	9,13%	**9,41**%

En los datos de la Tabla 2, para ambas formaciones, las facies de playa tienen una porosidad media mayor que la de las facies de aguas someras. Sin embargo, cuando se toman en conjunto las dos formaciones la relación se revierte y la porosidad de las facies de aguas someras resulta mayor que la de las facies de playa.

En este artículo se muestran varios ejemplos de la paradoja de Simpson en los datos geoquímicos de los yacimientos lateríticos del nordeste cubano y se discuten problemas relacionados con las condiciones de aparición y efectos de la misma.

2. ENFOQUE PROBABILÍSTICO DE LA PARADOJA DE SIMPSON

Sean $A = A_1$, $B = B_1$ y $C = C_1$ tres eventos aleatorios de un espacio muestral dado, $A_2 = \overline{A}$, $B_2 = \overline{B}$ y $A_2 = \bar{A}$, $B_2 = \bar{B}$ y $C_2 = \bar{C}$ sus respectivos complementos, y sea:

$P_{ijk} = P(A_i \bigcap B_j \bigcap C_k); i, j, k \in \{1, 2\}$

 ISSN 1993 8012

Como los eventos $A_i \cap B_j \cap C_k;\ i, j, k \in \{1,2\}$ son mutuamente excluyentes y su unión coincide con el espacio muestral, se cumple que:

$$\sum_{i,j,k} p_{ijk} = 1$$

La paradoja de Simpson ocurre cuando se cumplen las siguientes condiciones:

$$p_{111}p_{221} \geq p_{121}p_{211} \quad (1)$$

$$p_{112}p_{222} \geq p_{122}p_{212} \quad (2)$$

$$(p_{111} + p_{112})(p_{221} + p_{222}) \leq (p_{121} + p_{122})(p_{211} + p_{212}) \quad (3)$$

Donde, al menos una de las desigualdades es estricta o las tres desigualdades se sustituyen por su opuesto (ver, por ejemplo, Pavlides & Perlman 2009).

Si se le suma el término $p_{111}p_{121}$ a ambos miembros de la desigualdad (1) se obtiene:

$$p_{111}p_{221} + p_{111}p_{121} \geq p_{121}p_{211} + p_{111}p_{121}$$
$$p_{111}(p_{221} + p_{121}) \geq p_{121}(p_{211} + p_{111})$$

Si además se tiene en cuenta que:

$$p_{221} + p_{121} = P(\bar{B} \cap C),$$
$$p_{111} + p_{211} = P(B \cap C)$$

se obtiene que la desigualdad (1) es equivalente a:

$$P(A \cap B \cap C)P(\bar{B} \cap C) \geq P(A \cap \bar{B} \cap C)P(B \cap C)$$

que también se puede escribir en la forma: $P(A \mid B \cap C) \geq P(A \mid \bar{B} \cap C)$, donde, como es usual, mediante $P(X \mid Y)$ denotamos la probabilidad del evento X dado el evento Y (probabilidad condicional). Realizando transformaciones análogas en (2) obtenemos que dicha desigualdad es equivalente a $P(A \mid B \cap \bar{C}) \geq P(A \mid \bar{B} \cap \bar{C})$.

Por último, si se nota que:

$$p_{111} + p_{112} = P(A \cap B),$$
$$p_{221} + p_{222} = P(\bar{A} \cap \bar{B}),$$

 ISSN 1993 8012

$$p_{121} + p_{122} = P\left(A \cap \bar{B}\right),$$
$$p_{211} + p_{212} = P\left(\bar{A} \cap B\right),$$

se obtiene que la desigualdad (3) es equivalente a:

$$P\left(A \cap B\right) P\left(\bar{A} \cap \bar{B}\right) \leq P\left(A \cap \bar{B}\right) P\left(\bar{A} \cap B\right) \quad (4)$$

La expresión (4), a su vez, puede ser transformada si a cada término de la desigualdad se le suma $P(A \cap B)P(A \cap \bar{B})$convirtiéndola en:

$$P\left(A \cap B\right) P\left(\bar{B}\right) \leq P\left(A \cap \bar{B}\right) P\left(B\right)$$

De aquí, definitivamente, se desprende que la paradoja de Simpson ocurre si dados tres eventos A, B y C se cumplen las siguientes condiciones:

$$P(A \mid B \cap C) \geq P(A \mid \bar{B} \cap C) \quad (5)$$
$$P(A \mid B \cap C) \geq P(A \mid \bar{B} \cap C) \quad (6)$$
$$P(A \mid B) \leq P(A \mid \bar{B}) \quad (7)$$

Donde, al menos una de las desigualdades es estricta o las tres desigualdades se sustituyen por su opuesto.

Las desigualdades 5, 6 y 7 se pueden ver más fácilmente en la siguiente tabla de contingencia 2x2 (Tabla 3) en la cual se enfatizan con negritas los mayores valores de cada fila.

Tabla 3. Probabilidades condicionales del evento A dadas las combinaciones dos a dos de $B, C, \bar{B}, \bar{C}$.

	B	$\bar{B}$
C	$\boldsymbol{P\left(A \mid B \cap C\right)}$	$P\left(A \mid \bar{B} \cap C\right)$
$\bar{C}$	$\boldsymbol{P(A \mid B \cap \bar{C})}$	$P(A \mid \bar{B} \cap \bar{C})$
En conjunto	$P(A \mid B)$	$\boldsymbol{P(A \mid \bar{B})}$

Obsérvese que la Tabla 3 es el caso general de los ejemplos mostrados en las Tablas 1 y 2.

2.1. Condiciones y frecuencia de la aparición de la paradoja de Simpson

Aunque la paradoja de Simpson es conocida desde hace más de un siglo ha sido considerada muchas veces como un fenómeno anómalo y exótico. Recientemente, Pavlides & Perlman (2009) han demostrado

 ISSN 1993 8012

que no es exactamente así. De hecho, si se asume que las probabilidades $p_{ijk};\ i,j,k \in \{1,2\}$ están uniformemente distribuidas sobre el simplex:

$$\sum_{i,j,k} p_{ijk} = 1$$

Entonces, la probabilidad de que $\{p_{ijk}\}$ satisfaga las condiciones 1, 2 y 3 es de 1,66 % (Pavlides & Perlman 2009).

Veamos bajo qué condiciones la paradoja de Simpson ocurre. Supongamos, sin perder la generalidad que:

$$P(A|B\cap C) \geq P(A \mid B\cap \bar{C})$$

Si se tiene en cuenta que:

$$P(A|B) = P(A|B\cap C)P(C|B) + P(A|B\cap \bar{C})P(\bar{C}|B),$$

$$P(A|\bar{B}) = P(A|\bar{B}\cap C)P(C|\bar{B}) + P(A|\bar{B}\cap \bar{C})P(\bar{C} \mid \bar{B}),$$

se aprecia que $P(A|B)$ y $P(A|\bar{B})$ se expresan como combinaciones convexas de los valores $P(A|B\cap C)$, $P(A|B\cap \bar{C})$, $P(A|\bar{B}\cap C)$ y $P(A|\bar{B}\cap \bar{C})$. Los coeficientes de estas combinaciones convexas son las probabilidades condicionales $P(C|B)$, $P(\bar{C}|B)$, $P(C|\bar{B})$ y $P(\bar{C} \mid \bar{B})$. La convexidad se desprende de las identidades:

$$P(C|B) + P(\bar{C}|B) = 1$$

$$P(C|\bar{B}) + P(\bar{C}|\bar{B}) = 1$$

Es fácil ver entonces que para la aparición de la paradoja de Simpson es necesario que se cumpla la desigualdad múltiple:

$$P(A|B\cap C) \geq P(A|\bar{B}\cap C) \geq P(A|B\cap \bar{C}) \geq P(A|\bar{B}\cap \bar{C}), \qquad (8)$$

y además que:

$$P(C|B) < P(C \mid \bar{B}) \qquad (9)$$

La condición (9) es más general que la expresada en Ma (2009) ya que no asume que el espacio muestral sea finito. Si se supone que el espacio muestral al que pertenecen los eventos A, B y C es finito y

 ISSN 1993 8012

denotamos mediante M_1 , M_2 , N_1 y N_2 la cantidad de elementos (o la cantidad de muestras, si se prefiere) de los conjuntos $C \cap B$, $\bar{C} \cap B$, $C \cap \bar{B}$ y $\bar{C} \cap \bar{B}$, respectivamente, entonces se tiene en ese caso:

$$P(C|B) = \frac{M_1}{M_1 + M_2}; \qquad P(C|\bar{B}) = \frac{N_1}{N_1 + N_2}$$

y por tanto la desigualdad (9) se reduce a:

$$M_1(N_1 + N_2) < N_1(M_1 + M_2) \Leftrightarrow M_1 N_2 < N_1 M_2 \Leftrightarrow \frac{M_1}{M_2} < \frac{N_1}{N_2} \qquad (10)$$

Esta última condición es la reflejada por Ma (2009).

3. RESULTADOS Y DISCUSIÓN

En este apartado se muestran tres ejemplos de la paradoja de Simpson en yacimientos lateríticos del norte oriental de Cuba. En todos los casos se comparan regiones rectangulares disjuntas de un área no menor de doce hectáreas. Los resultados geoquímicos que se exponen corresponden a muestras tomadas metro a metro en una red de pozos de exploración.

En la Tabla 4 se compara el contenido de hierro en dos zonas desagregándolo en dos de las capas litológicas de mayor importancia para la industria del níquel.

Tabla 4. Contenido promedio de hierro (y número de muestras) en dos capas litológicas de dos bloques rectangulares en un yacimiento de níquel del oriente de Cuba

Litología	Zona A	Zona B
Ocres Estructurales Finales	**44,74**% (107)	43,14% (138)
Ocres Estructurales Iniciales	**28,33**% (206)	28,02% (125)
En ambas capas litológicas	33,94%	**35,95**%

Puede apreciarse que la desigualdad presente en cada capa se revierte cuando se consideran ambas en conjunto. Si denotamos mediante B el evento consistente en que una muestra tomada al azar pertenezca a la zona A y mediante C el evento en que una muestra tomada al azar sea clasificada como de ocres estructurales finales, representaríamos la

 ISSN 1993 8012

Tabla 4 como un caso particular de la Tabla 3. Para este caso las probabilidades condicionales presentes en (9) son:

$$P(C|B)=\frac{107}{107+206}=0,34; \qquad P(C|\bar{B})=\frac{138}{138+125}=0,52$$

por lo que la condición necesaria (9) se cumple. En la Tabla 5 se muestra una inversión del contenido de sílice.

Tabla 5. Contenido promedio de SiO_2 en dos capas litológicas de dos bloques rectangulares en un yacimiento de níquel del oriente de Cuba

Litología	Bloque E	Bloque F
Ocres estructurales Iniciales	**24,39** % (110)	22,69 % (140)
Ocres estructurales Finales	**8,74** % (193)	6,57 % (127)
En ambas capas litológicas	14,42 %	**15,02** %

En el caso de la Tabla 5 las probabilidades condicionales de la expresión (9) son:

$$P(C|B)=\frac{110}{110+193}=0,36; \qquad P(C|\bar{B})=\frac{140}{140+127}=0,52$$

Es fácil ver que en las Tablas 4 y 5 la causa esencial de la aparición de la paradoja es la desproporcionalidad de las muestras en las subpoblaciones, derivada del diferente espesor de cada litología en las zonas analizadas. Por otra parte, la influencia de las variaciones de espesor es más relevante en las concentraciones de Fe y SiO_2 que en la zona de donde fue tomada la muestra.

La paradoja de Simpson deja abierto el dilema de cuáles datos usar para la toma de decisiones, los datos particionados o los agregados. En el ejemplo del tratamiento para cálculos renales si se desea tratar a un paciente con un cálculo cuyo tamaño aún es desconocido ¿cuál tratamiento debería recomendarse con base en la Tabla 1?.

Pudiera parecer que siempre es mejor usar los datos particionados, sin embargo, no siempre es así. Pearl (2000) muestra que en muchas ocasiones son los datos agregados y no los particionados los que ofrecen la elección correcta. En todo caso podemos disminuir mucho la ambigüedad si se definen correctamente qué variables influyen en el indicador analizado, estableciendo las relaciones causales entre ellas y representando estas relaciones en redes causales bayesianas (Pearl 2000).

 ISSN 1993 8012

La Tabla 6 muestra una reversión del contenido de magnesio en tres capas litológicas a la vez.

Tabla 6. Contenido promedio de magnesio (y número de muestras) en tres capas litológicas de dos bloques rectangulares en un yacimiento de níquel del oriente de Cuba

Litología	Bloque C	Bloque D
Ocres Estructurales Iniciales	**7,98** % (241)	7,94 % (208)
Ocres Estructurales Finales	**2,93** % (267)	1,93 % (70)
Ocres Inestructurales sin perdigones	**2,90** % (87)	1,63 % (69)
En las tres capas litológicas	4,97 %	**5,47** %

Debemos notar que la probabilidad de que aparezca una reversión de Simpson en una tabla de contingencia de 3x2, asumiendo la distribución uniforme de las probabilidades $p_{ijk}; i, j, k \in \{1,2\}$ es de 0,57% (Pavlides & Perlman 2009), por lo que no deja de ser muy interesante que se hayan encontrado estas reversiones en los datos de las exploraciones de yacimientos lateríticos del oriente cubano.

Los efectos de la paradoja de Simpson son mayores de los que pudiera parecer a primera vista, sobre todo en virtud de que el análisis de los datos queda, en proporción cada vez más creciente, a cargo de sistemas automatizados.

Las herramientas para coleccionar datos en mapas digitales, los datos satelitales y la difusión global de los Sistemas de Información Geográfica (GIS) han generado conjuntos cada vez más grandes de datos. Por ejemplo, el sistema de observación de la Tierra de la NASA (NASA 2006) ofrece observaciones de la superficie terrestre, de la biosfera, de la atmósfera y del océano en el orden de terabytes de datos diarios. La detección automatizada de patrones regionales, a partir de esta enorme información disponible, es uno de los retos que tiene la minería de datos, en especial la detección de aquellos patrones de interés que no están explícitamente representados en las bases de datos geo-referenciadas.

Los sistemas de minería de datos existentes generalmente se enfocan en la detección de patrones globales y fallan en la capacidad para detectar patrones locales. Celepcikay *et al* (2009) y Ding *et al* (2011) reportan ejemplos en los que la imposibilidad de hacer conclusiones regionales adecuadas tiene como origen la paradoja de Simpson.

4. CONCLUSIONES

Se reportan tres casos de ocurrencia de la paradoja de Simpson en la exploración de yacimientos lateríticos cubanos. En los tres casos la causa de la aparición de este fenómeno es esencialmente provocada porque los contenidos de los elementos analizados dependen en mayor grado de los espesores de cada litología que de su pertenencia a alguna zona específica del yacimiento. No se excluye, sin embargo, que en otras zonas y con otras tablas de contingencia, se encuentren resultados similares. La única manera de garantizar que no aparezcan las reversiones de Simpson en tablas de contingencia y, consecuentemente, evitar el sesgo en la inferencia estadística que ella provoca, es tratar de que la elección de las subpoblaciones se haga de manera que no se cumplan las condiciones necesarias para su aparición, en particular garantizar que, bajo las hipótesis planteadas anteriormente, se incumpla la desigualdad (8) o se satisfaga la condición $P(C|B) \geq P(C \mid \overline{B})$.

5. REFERENCIAS

CELEPCIKAY, O. U.; EICK, C. F; ORDONEZ, C. 2009: Regional Pattern Discovery in Geo-Referenced Datasets Using PCA. *Lecture Notes In Artificial Intelligence,* 5632.

CHARIG, C. R.; WEBB, D. R.; PAYNE, S. R.; WICKHAM O. E. 1986: Comparison of Treatment of Renal Calculi by Open Surgery, Percutaneous Nephrolithotomy and Extracorporeal Shockwave Lithotripsy. *British Medical Journal (Clin. Res. Ed.),* 292 (6524): 879–882.

DING, W.; EICK, C. F; WANG, J.; YUAN, X. J. 2011: A framework for regional association rule mining and scoping in spatial datasets. *Geoinformática* 15(1):1–28.

JULIOUS, S. A.; MULLEE, M. A. 1994: Confounding and Simpson's paradox. *British Medical Journal* 209 (6967): 1480–1481.

MA, Y. Z. 2009: Simpson's Paradox in Natural Resource Evaluation. *Math Geosci* 41: 193–213.

MA, Y. Z.; MA, A. M. 2011: Simpson's Paradox and Other Reversals in Basketball: Examples from 2011 NBA Playoffs. *International Journal of Sports Science and Engineering* 05(03):145-154.

NASA 2006: NASA's Earth Observing System Project. Disponible en http://eospso.gsfc.nasa.gov/.

PAVLIDES, M. G; PERLMAN, M. D. 2009: How Likely Is Simpson's Paradox? *American Statistician* 63(3):226-233.

PEARL, J. 2000: Causality: Models, Reasoning, and Inference. *Cambridge University Press*. ISBN 0-521-77362-8.

TERWILLIGER, S.; SCHIELD, M. 2004: Frequency of Simpson's Paradox in NAEP data. Presented at AERA, September.

ZUZOVSKY, R.; STEINBERG D. M. 2011:Achievement data in IEA studies and Simpson's Paradox. *Studies In Educational Evaluation* 37(2-3): 141-151.

Rafael Arturo Trujillo-Codorníu rtrujillo@ismm.edu.cu
Doctor en Ciencias Matemáticas. Profesor Titular.
Departamento de Matemáticas e Informática.
Instituto Superior Minero Metalúrgico, Moa, Cuba.

Santiago Bernal-Hernández sbernal@ismm.edu.cu
Doctor en Ciencias Técnicas. Profesor Titular.
Departamento de Minas.
Instituto Superior Minero Metalúrgico, Moa, Cuba.

Mirelis Rasúa-López mrasua@ismm.edu.cu
Licenciada en Matemáticas. Profesora Asistente.
Departamento de Matemáticas e Informática.
Instituto Superior Minero Metalúrgico, Moa, Cuba.