MODELACIÓN DE DATOS ESPACIALES CENSURADOS

José Elías Rodríguez Muñoz

Comunicación Técnica No I-02-23/24-10-2002
(PE/CIMAT)

Modelación de Datos Espaciales Censurados

José Elías Rodríguez Muñoz

Facultad de Matemáticas

Universidad de Guanajuato

México

Contenido

1 Introducción 1

2 Tipos de censura 5

3 Censura y modelos con variables aleatorias independientes 9

3.1 Censura Tipo II 9

3.2 Censura Tipo II progresiva 12

3.3 Censura Tipo I 16

4 Censura y variables latentes 22

5 Datos espaciales 27

6 Comentarios …nales 42

1 Introducción

Los instrumentos de medición que se utilizan para estudiar la naturaleza tienen una precisión[1] ...nita. Esto no implica que las observaciones que hagamos de algún fenómeno estén todas dentro de tal precisión. Por lo que es de esperar que en los datos registrados encontremos valores dentro de la precisión del instrumento de medición y otros fuera, cuyos valores no los conoceremos. Este tipo de datos, así registrados, se denominan datos censurados. También cuando experimentamos es posible encontrarse con datos censurados. Por ejemplo, si estamos probando la vida útil de 20 aparatos eléctricos y decidimos parar el experimento cuando ocurran las primeras 15 fallas, los restantes 5 valores de la vida útil de los aparatos serán desconocidos y sólo tendremos registrado que son mayores que la décima quinta falla registrada.

En datos espaciales (datos georeferenciados) también se pueden encontrar datos censurados. Por ejemplo en estudios de mineria, de monitoreo ambiental de contaminantes o de precipitación pluvial se pueden encontrar datos por debajo del límite de detección[2] del procedimiento de medición. Estos datos así registrados se consideran como datos censurados.

[1] Por precisión se entenderá el conjunto de posibles valores que puede generar un procedimiento de medición.

[2] Por límite de detección se entenderá un valor que acota la precisión.

El presente trabajo tiene como objetivo mostrar como se modelan fenómenos o experimentos que producen datos con censura, en particular datos censurados con una estructura de dependencia espacial.

Para entender mejor este tipo de datos, sea $z_n = \{z_{1n}, \ldots, z_{nn}\}$ un conjunto de datos de algún experimento o fenómeno. Para el propósito del presente trabajo se supondrá que los datos son números reales y producto de mediciones.

El conjunto z_n consta de **datos completos** si tiene registrado el valor de todos los elementos que lo componen. El conjunto z_n tiene **datos perdidos** si el registro de al menos uno de los datos se ha extraviado. El conjunto consta de **datos truncados** si sus valores, todos conocidos, pertenecen a un subconjunto predeterminado del conjunto de posibles valores de experimento o fenómeno de interes; por ejemplo todos los valores son menores que un predeterminado número real. El conjunto z_n tiene **datos censurados** si para al menos uno de los datos se desconoce su valor, pero si se conoce un subconjunto de los números reales que contiene tal valor. A este subconjunto donde pertenece el verdadero valor del dato censurado se le denominará **conjunto de censura**.

Los datos censurados aparecen frecuentemente cuando el experimento o

fenómeno bajo estudio involucra el tiempo de vida o supervivencia. En pruebas de vida útil conducidas en Física o Ingeniería, la censura es frecuentemente inducida para acortar el tiempo de prueba. El libro de [Nelson, 1982] presenta un compendio de métodos para analizar datos censurados que surgen en aplicaciones de Ingeniería.

Mientras que para tiempos de vida o supervivencia se tiene una basta literatura para modelar y analizar datos censurados, existen pocos métodos formales para modelar y analizar datos espaciales censurados. Recientemente, en el trabajo de [Stein, 1992] presenta métodos monte carlo para calcular las distribuciones condicionales de un campo aleatorio y en [Militino y Ugarte, 1999] desarrollan una metodología basada en el algoritmo EM para el análisis de datos espaciales censurados.

Para entender mejor como se modelan los datos espaciales censurados, primero necesitamos saber como se clasi...ca la censura de acuerdo al mecanismo que la produce y la clasi...cación de acuerdo a los tipos de conjuntos de censura; la sección 2 muestra estas clasi...caciones. En segundo lugar es básico conocer la forma en que se modelan fenómenos o experimentos con variables aleatorias independientes y bajo censura, esto se muestra en la sección 3. Existe una relación entre los modelos para el estudio de datos censurados y las

variables latentes, ésta se presenta en la sección 4. Por último, en la sección 5 se presenta como se modelan y analizan datos espaciales censurados.

2 Tipos de censura

Existen varios tipos de datos censurados según sea el mecanismo que produce la censura y el tipo de conjunto que contiene el valor del dato censurado. Una clasi…cación similar a la que se presenta aquí se puede encontrar en la sección 1.3 de [Cohen, 1991]. También una relación entre los procesos de medición y los datos etiquetados como censurados la podemos encontrar en el trabajo de [Lambert et al., 1991].

De acuerdo al mecanismo de censura, los conjuntos de datos censurados son clasi…cados como censurados del Tipo I y censurados del Tipo II. Un conjunto de datos es censurado del Tipo I si los conjuntos de censura son conocidos de antemano a la observación de los datos; pero el número de valores censurados, denotado por r, no se conoce. En contraposición, un conjunto de datos es censurado del Tipo II si el número de valores censurados es …jado antes de su registro, pero los conjuntos de censura no lo son. En datos espaciales la censura que se observa es la de Tipo I.

Ahora, de acuerdo a los conjuntos de censura, los datos censurados son clasi…cados como sigue.

1. Datos con censura simple por la izquierda: para cada observación

censurada z, se conoce sólo que $z < T$, mientras que para cada observación medida, $z \geq T$. En datos censurados del Tipo I, T es un punto …jo de censura. En datos censurados del Tipo II, $T = z_{(r+1):n}$, esto es, el estadístico de orden $r + 1$ en un conjunto de tamaño n.

2. Datos con censura simple por la derecha: para cada observación censurada z, se conoce sólo que $z > T$; mientras que para cada observación medida, $z \leq T$. En datos censurados del Tipo I, T es un punto …jo de censura. En datos censurados del Tipo II, $T = z_{(n-r):n}$.

3. Datos con censura doble: para cada observación medida z, se conoce su valor que satisface $T_1 \leq z \leq T_2$. Además, existen r_1 observaciones censuradas por la izquierda y r_2 por la derecha, $r_1 + r_2 = r$, de las cuales solo se conoce que $z < T_1$ o $z > T_2$ según sea el caso. En datos censurados del Tipo I, T_1 y T_2 son puntos …jos de censura. En datos censurados del Tipo II, $T_1 = z_{(r_1+1):n}$ y $T_2 = z_{(n-r_2):n}$. Es inmediato generalizar este tipo de censura para establecer lo que se conoce como censura multiple.

Es conveniente hacer notar que un conjunto de censura tiene que ser un Boreleano de los números reales. Por ejemplo, si la censura es por la

izquierda con punto de censura T, entonces el conjunto de censura es de la forma $(-1;T)$. Esto nos muestra que la clasi...cación de los datos censurados de acuerdo a los conjuntos de censura mostrada arriba, enlista un conjunto pequeño de posibilidades. Sin embargo, estos tipos de datos censurados en la clasi...cación son los más usados en la práctica (ver por ejemplo el libro de [Nelson, 1982]), aunque existen teóricamente muchas otras posibilidades.

Existe otro tipo de datos censurados que se asemeja más en la clasi...cación de la censura por su mecanismo que la produce, pero por su naturaleza la presento aparte y se estudia con más detalle en la sección 3.2. Este tipo de censura se dá en mediciones de tiempo de vida o reacción de elementos bajo observación o experimentación. Por esto, su descripción se referirá a estas mediciones de tiempo. Un conjunto de datos es censurado del Tipo II progresivo si los primeros n_1 tiempos son registrados, entonces r_1 de los restantes $n - n_1$ elementos son removidos de la observación o del experimento, dejando $n - n_1 - r_1$ elementos aún presentes. Cuando a n_2 elementos de los restantes se les registra su tiempo de vida o reacción, entoces se retirán r_2 elementos del resto. Este proceso continua hasta completar n datos entre tiempos registrados y datos censurados (removidos).

Debido al procedimiento de medición, el análisis de datos espaciales sólo

contempla la censura simple por la izquierda o por la derecha del Tipo I, como se verá en la sección 5. Sin embargo, en este trabajo se presentan otros tipos de censura con el objetivo de tener una percepción más amplia de los diferentes tipos de datos censurados.

3 Censura y modelos con variables aleatorias independientes

Los datos censurados aparecen frecuentemente cuando la variable de interés es el tiempo de vida o supervivencia o el tiempo de reacción de los objetos o individuos bajo prueba u observación. Para estos casos si las pruebas u observaciones sobre los objetos o individuos se realizan independientemente unas de otras, las variables aleatorias continuas, digamos $Z_n = \{Z_{1n}; \ldots Z_{nn}\}$, que modelan el comportamiento del experimento o fenómeno bajo estudio se suponen independientes e idénticamente distribuidas. Para la presente sección siempre se considerarán modelos de este estilo.

3.1 Censura Tipo II

Como ya se mencionó en la sección 2, el conjunto de datos z_n es censurado del Tipo II si el número de valores censurados r, $1 \quad r \quad n$, es ...jado antes de su registro, pero los conjuntos de censura no. Por ejemplo, este tipo de censura se dá cuando n elementos se ponen a prueba pero en lugar de continuar con el experimento hasta que todos hayan fallado, la prueba termina en el momento que se registra la $(n - r)$-ésima falla. Este número r es ...jado antes

de comenzar con las pruebas. Además este tipo de experimentos se utilizan para ahorrar tiempo y dinero, ya que puede pasar mucho tiempo antes de observar que todos los elementos bajo prueba fallen.

Ilustraré la modelación de datos censurados del tipo II para el caso de la censura por la derecha. Una exposición similar se puede encontrar en la sección 1.4.1 de [Lawless, 1982]. La extensión a otros tipos de conjuntos de censura es casi inmediata.

Formalmente, si las variables aleatorias Z_n modelan el experimento bajo estudio, las observaciones resultantes del mecanismo de censura del Tipo II y por la derecha se modelan con los estadísticos de orden $Z_{1:n}; : : : ; Z_{n ¡ r:n}$. Si Z_n es un conjunto de n variableas aleatorias continuas, independientes e idénticamente distribudidas con Función de Distribución G, entonces la función de densidad conjunta resultante de $Z_{1:n}; : : : ; Z_{n ¡ r:n}$ es

$$\frac{n!}{r!}\left(\prod_{k=1}^{n-r} g(z_{k:n})\right)(1 ¡ G(z_{n ¡ r:n}))^r; \qquad (3.1)$$

para $z_{1:n} < ¢¢¢ < z_{n ¡ r:n}$. Aquí g es la función de densidad de G y $z_{k:n}$ se utiliza para representar el k-ésimo dato observado ordenado. Es conveniente mencionar que en el contexto de estudios de tiempos de vida, la función

$1 - G$ (parte de la expresión (3:1)) se denomina la función de supervivencia o función de con...abilidad.

Para cualquier modelo paramétrico la inferencia se puede basar en la función de densidad conjunta (3:1), la cual proporcionará la función de verosimilitud y de ésta podemos derivar las propiedades estadísticas de los procedimientos utilizados para la inferencia.

Ilustraré lo anterior con el siguiente ejemplo.

EJEMPLO 3.1 Supongamos que las variables aleatorias Z_n tienen Función de Distribución asociada Exponencial (γ), cuya función de densidad es de la forma

$$g(z) = \begin{cases} 0 & \text{si } z < 0 \\ \frac{1}{\gamma}\exp\left(-\frac{z}{\gamma}\right) & \text{en otro caso.} \end{cases}$$

Así, la función de densidad conjunta de $Z_{1:n}, \ldots, Z_{n-r:n}$ resulta ser

$$\begin{aligned} &\frac{n!}{r!}\left(\frac{1}{\gamma^{n-r}}\right)\exp\left(-\frac{1}{\gamma}\sum_{k=1}^{n-r} z_{k:n}\right)\exp\left(-r\frac{z_{n-r:n}}{\gamma}\right) \\ &= \frac{n!}{r!}\left(\frac{1}{\gamma^{n-r}}\right)\exp\left(-\frac{1}{\gamma}\left[\sum_{k=1}^{n-r} z_{k:n} + r z_{n-r:n}\right]\right); \end{aligned} \tag{3.2}$$

para $z_{1:n} < \cdots < z_{n-r:n}$ y cero en otro caso.

Entre los instrumentos para hacer inferencia podemos contar con el estadístico

$$T = \sum_{k=1}^{n-r} Z_{k:n} + rZ_{n-r:n};$$

que es un estadístico su…ciente para el parámetro ¾: Además, el estimador de máxima verosimilitud de este mismo parámetro es $T=(n-r)$.

3.2 Censura Tipo II progresiva

Recordemos de la sección 2 que en este tipo de censura los primeros n_1 tiempos son registrados, después r_1 de los restantes $n-n_1$ elementos son removidos de la observación o del experimento, dejando $n-n_1-r_1$ elementos aún presentes. Cuando a n_2 elementos de los restantes se les registra su tiempo de vida o reacción, entoces se retirán r_2 elementos del resto. Este proceso continuará hasta completar n datos entre tiempos registrados y datos censurados.

Para ejempli…car la función de densidad resultante de este tipo de censura sólo consideraremos la censura en dos etapas. Esto es, después de registrar los primeros n_1 tiempos de falla se retiran r_1 elementos, de los restantes

$n - n_1$. El experimento termina cuando los siguientes n_2 tiempos de vida se registran. En este momento existen todavía $n - n_1 - r_1 - n_2$ elementos sin fallar.

La exposición que sigue es similar a la mostrada en la sección 1.4.1 de [Lawless, 1982].

Las primeras n_1 observaciones, de la primera etapa de la censura, se modelan con los estadísticos de orden $Z_{1:n}, \ldots, Z_{n_1:n}$ cuya función de densidad está dada por la expresión (3:1); con $n - r = n_1$. Las n_2 observaciones registradas después de observar las primeras n_1 y retirar r_1 elementos del estudio se modelan con las variables aletorias $Z^{*}_{1:(n-n_1-r_1)}, \ldots, Z^{*}_{n_2:(n-n_1-r_1)}$ que son los primeros n_2 estadísticos de orden de las nuevas variables aleatorias independientes e idénticamente distribuidas $Z^{*}_{1(n-n_1-r_1)}, \ldots, Z^{*}_{(n-n_1-r_1)(n-n_1-r_1)}$. Estas últimas variables aleatorias se obtienen asignandoles aleatoriamente los valores restantes de la primera etapa de censura.

La función de densidad condicional marginal de cada una de estas últimas variables es

$$g^{*}(z \mid z_{n_1:n}) = \begin{cases} \dfrac{g(z)}{1 - G(z_{n_1:n})} & \text{si } z > z_{n_1:n} \\ 0 & \text{en otro caso,} \end{cases} \qquad (3.3)$$

donde G es la Función de Distribución original y g es su respectiva función de densidad. Además $z_{n_1:n}$ representa el valor observado del estadístico de orden $Z_{n_1:n}$. En verdad la función de densidad de la expresión (3:3) es la función de densidad marginal de cada una de las variables aleatorias

$$\left(Z^{\star}_{1(n-n_1-r_1)},\ldots;Z^{\star}_{(n-n_1-r_1)(n-n_1-r_1)}\right)$$

dada las observaciones $\{z_{1:n},\ldots;z_{n_1:n}\}$, pero la expresión resultante de este condicionamiento es como se expresa en la ecuación (3:3):

Así, la función de densidad conjunta condicional de

$$Z^{\star}_{1:(n-n_1-r_1)},\ldots;Z^{\star}_{n_2:(n-n_1-r_1)}$$

estará dada por la expresión

$$g_2\left(z^{\star}_{1:(n-n_1-r_1)},\ldots;z^{\star}_{n_2:(n-n_1-r_1)}\,\middle|\,z_{n_1:n}\right) = \frac{(n-n_1-r_1)!}{(n-n_1-r_1-n_2)!}\left(\prod_{k=1}^{n_2} g\left(z^{\star}_{k:(n-n_1-r_1)}\right)\right)\left(\frac{1-G\left(z^{\star}_{n_2:(n-n_1-r_1)}\right)}{1-G(z_{n_1:n})}\right)^{(n-n_1-r_1-n_2)}$$

$$= \frac{(n - n_1 - r_1)!}{(n - n_1 - r_1 - n_2)!} \cdot \frac{1}{(1 - G(z_{n_1:n}))^{(n - n_1 - r_1)}} \cdot \left[\prod_{k=1}^{n_2} g\left(z^{*}_{k:(n - n_1 - r_1)}\right) \cdot \left(1 - G\left(z^{*}_{n_2:(n - n_1 - r_1)}\right)\right)^{(n - n_1 - r_1 - n_2)} \right].$$

Ahora, la función de densidad conjunta de

$$Z_{1:n}, \ldots, Z_{n_1:n} \text{ y } Z^{*}_{1:(n - n_1 - r_1)}, \ldots, Z^{*}_{n_2:(n - n_1 - r_1)},$$

que modela las observaciones registradas y censuradas, está dada por

$$g_1(z_{1:n}, \ldots, z_{n_1:n})\, g_2\left(z^{*}_{1:(n - n_1 - r_1)}, \ldots, z^{*}_{n_2:(n - n_1 - r_1)} \mid z_{n_1:n}\right), \tag{3.4}$$

donde g_1 es la función de densidad conjunta de $Z_{1:n}, \ldots, Z_{n_1:n}$ y dada por la expresión (3:1); con $n - r = n_1$. Después de algunas manipulaciones algebraicas la función de densidad (3:4) queda como

$$\frac{n!\,(n - n_1 - r_1)!}{(n - n_1)!\,(n - n_1 - r_1 - n_2)!} \cdot \left[\prod_{k=1}^{n_1} g(z_{k:n}) \cdot \left(1 - G(z_{n_1:n})\right)^{r_1} \right] \cdot \left[\prod_{k=1}^{n_2} g\left(z^{*}_{k:(n - n_1 - r_1)}\right) \cdot \left(1 - G\left(z^{*}_{n_2:(n - n_1 - r_1)}\right)\right)^{(n - n_1 - r_1 - n_2)} \right];$$

donde $z_{1:n} < ¢¢¢ < z_{n_1:n} < z^{¤}_{1:(n\,¡\,n_1\,¡\,r_1)} < ¢¢¢ < z^{¤}_{n_2:(n\,¡\,n_1\,¡\,r_1)}$.

En forma semejante podemos obtener la función de densidad para el caso de más de dos etapas de censura, aunque los calculos pueden complicarse rápidamente.

3.3 Censura Tipo I

Recordemos que un conjunto de datos es censurado del Tipo I si los conjuntos de censura son conocidos de antemano pero no el número de datos censurados.

Para este tipo de censura es conveniente reexpresar nuestros datos z_n como $z_n = f(z_{1n}; ±_{1n}); : : : ; (z_{nn}; ±_{nn})g$ con conjuntos de censura $fB_{1n}; : : : ; B_{nn}g$ donde

$$±_{jn} = \begin{cases} 1 & \text{si } z_{jn} \,2\, B^{c}_{jn} \\ 0 & \text{si } z_{jn} \,2\, B_{jn}; \end{cases}$$

para $j = 1; : : : n$. Esto es, $±_{jn}$ nos indica si la observación j se tiene registrada ($±_{jn} = 1$) o censurada ($±_{jn} = 0$). También, las variables aleatorias Z_n que modelan este tipo de datos las presentamos como $Z_n = f(Z_{1n}; ¢_{1n}); : : : ; (Z_{nn}; ¢_{nn})g$, donde las variables Zs modelan el comportamiento del experimento o fenómeno bajo estudio y las variables $¢$s modelan el mecanismo de censura.

Si el dato es de la forma $(z; 1)$ quiere decir que se tiene registrado el valor de z y que no pertenece al conjuto de censura, digamos B. Además, la contribución marginal a la verosimilitud de este dato es

$$Pr(z - \varepsilon < Z \leq z + \varepsilon; \delta = 1) = Pr(z - \varepsilon < Z \leq z + \varepsilon; Z \in B^c).$$

Para ε su…cientemente pequeño se puede lograr que $[z - \varepsilon; z + \varepsilon] \subset B^c$ para obtener que

$$Pr(z - \varepsilon < Z \leq z + \varepsilon; \delta = 1) = Pr(z - \varepsilon < Z \leq z + \varepsilon).$$

La parte derecha de la igualdad es proporcional a $2\varepsilon g(z)$, donde g es la función de densidad de la variable aleatoria Z.

Ahora, si el dato es de la forma $(z; 0)$, entonces no se tiene registrado el valor de z pero si se sabe que pertenece al conjunto de censura B. Así, la contribución marginal a la verosimilitud está dada por $Pr(Z \in B)$.

De esta forma, si las variables aleatorias $\{Z_{1n}, \ldots, Z_{nn}\}$ son independientes e idénticamente distribuidas con función de densidad asociada g, la

verosimilitud en los datos z_n la podemos expresar como

$$\prod_{k=1}^{n} [g(z_{kn})]^{\pm_{kn}} [Pr(Z_{kn} \in B_{kn})]^{1-\pm_{kn}} :$$

Una deducción similar de esta última expresión la podemos encontrar en [Kalb‡eisch y Prentice, 1980].

Ilustremos lo anterior con el siguiente ejemplo.

EJEMPLO 3.2 Supongamos que las variables aleatorias $\{Z_{1n}, \ldots, Z_{nn}\}$ tienen Función de Distribución Exponencial (γ), cuya función de densidad es de la forma dada en el ejemplo 3.1 y tiene censura simple por la derecha con punto de censura igual a z_d: Así, $B_{kn} = (z_d, +1)$; $k = 1, \ldots, n$; y la verosimilitud es

$$\prod_{k=1}^{n} \left[\frac{1}{\gamma}\exp\left(-\frac{z_{kn}}{\gamma}\right)\right]^{\pm_{kn}} \left[\exp\left(-\frac{z_d}{\gamma}\right)\right]^{(1-\pm_{kn})} : \qquad (3.5)$$

Dado el valor de la suma $\sum_{k=1}^{n} \pm_{kn}$, la expresión (3:5) es proporcional a la presentada en la ecuación (3:2), con $n - r = \sum_{k=1}^{n} \pm_{kn}$. El estimador de

máxima verosimilitud de ¾ es

$$\frac{\sum_{fk:\pm_{k:n}=1g} Z_{kn} + rz_d}{n \; ¡ \; r};$$

si $r > 0$: Notemos que algebraicamente es igual al estimador del ejemplo 3.1. Sin embargo, aquí r es aleatorio. Por tanto las propiedades estadísticas de este estimador cambian radicalmente, comparadas con las del estimador del ejemplo mencionado.

Una adición a la modelación de datos censurados (en general) es la aplicación de una transformación a los datos como parte del proceso de ésta. La …nalidad de la transformación es facilitar la identi…cación de un modelo para los datos transformados. La di…cultad posterior es estimar la característica del modelo que interesa en la escala original de los datos; además de identi…car un transformación adecuada.

Recordemos que los datos censurados de Tipo I se modelan originalmente con el conjunto de variables aleatorias $Z_n = f(Z_{1n}; ¢_{1n}); : : : ; (Z_{nn}; ¢_{nn})g$: Después de aplicar una transformación medible h (generalmente monótona) a los datos, las variables aleatorias que modelan los datos transformados serán $h(Z_n) = f(h(Z_{1n}); ¢_{1n}); : : : ; (h(Z_{nn}); ¢_{nn})g$: Para el conjunto de

datos $h(z_n)$ se espera que la identi…cación de un modelo estocástico sea más factible que para el conjunto original z_n.

El trabajo de [Shumway et al., 1989] servirá para ilustrar la aplicación de una transformación a datos censurados. En este artículo analizan datos censurados del Tipo I por la izquierda. La transformación aplicada a los datos es la de Box y Cox y de…nida por

$$h_{\lambda}(z_{kn}) = \begin{cases} \dfrac{z_{kn}^{\lambda} - 1}{\lambda} & \text{para } \lambda \neq 0 \\ \ln(z_{kn}) & \text{para } \lambda = 0; \end{cases}$$

donde z_{kn} es el valor del dato ($\pm_{kn} = 1$) o el punto de censura si corresponde a un dato censurado ($\pm_{kn} = 0$). Recordemos que esta transformación es útil cuando los datos transformados se modelan adecuadamente con una normal.

De esta forma en el artículo de [Shumway et al., 1989] el logaritmo natural de la verosimilitud resultante es

$$\sum_{k=1}^{n} \left\{ \pm_{kn} \left[-\frac{1}{2} \ln\left(2\pi\sigma^2\right) + (\lambda - 1)\ln(z_{kn}) - \frac{\left(\dfrac{z_{kn}^{\lambda} - 1}{\lambda} - \mu\right)^2}{2\sigma^2} \right] + [1 - \pm_{kn}] \ln\left[1 - \Phi\left(\frac{\dfrac{z_{kn}^{\lambda} - 1}{\lambda} - \mu}{\sigma} \right) \right] \right\}.$$

Utilizando el logaritmo de la verosimilitud anterior y el algoritmo EM en el proceso de estimación (ver la sección 5) los autores del mencionado artículo estiman los parámetros ^{1} y $¾^2$, ...jando primero el parámetro de la transformación ¸: Posteriormente, repitiendo lo anterior para varios valores del parámetro ¸ se encuentran los estimadores ...nales de los parámetros $(^1; ¾^2; ¸)$. Las estimaciones resultantes se utilizan por ejemplo en el cálculo de un intervalo de con...anza para la media del modelo en la escala original.

Otro ejemplo de modelación de datos censurados de Tipo I lo encontraremos en la sección 5.

Un método de modelación relacionado con los modelos para datos censurados es el que tiene que ver con variables latentes, que presento en la siguiente sección.

4 Censura y variables latentes

Los datos de ciertos tipos de fenómenos o experimentos tienen algunos de sus valores concentrados en una cantidad. Por facilidad de exposición supondré que esta cantidad de concentración es 0 y que el resto tienen valores positivos. Esto es, en los datos $z_n = \{z_{1n}; \ldots; z_{nn}\}$ algunos de sus valores son cero y el resto son positivos.

Un ejemplo de este tipo de datos es el monto o porcentaje del ingreso destinado al ahorro, encontrado en estudios de bienestar económico de individuos en una población. En este caso es común observar individuos que no destinan parte de sus ingresos al ahorro. En datos espaciales encontramos información sobre precipitación pluvial donde los ceros corresponden a periodos de ausencia de lluvia.

Las propiedades estadísticas de las variables aleatorias no negativas $Z_n = \{Z_{1n}; \ldots; Z_{nn}\}$; que modelan el comportamiento del fenómeno o experimento bajo estudio, quedan determinadas por las variables aleatorias latentes $Z_n^{\alpha} = \{Z_{1n}^{\alpha}; \ldots; Z_{nn}^{\alpha}\}$ y una función invertible h. Las mencionadas propiedades estadísticas de las variables aleatorias Z_n quedan especi…cadas

por la siguiente relación

$$Z_{kn} = \begin{cases} h(Z^{\alpha}_{kn}) & \text{si} \quad Z^{\alpha}_{kn} > 0 \\ 0 & \text{si} \quad Z^{\alpha}_{kn} \quad 0; \end{cases} \tag{4.1}$$

para $k = 1; : : : ; n$.

De esta forma, el conjunto de datos observados z_n produce un nuevo conjunto de datos, no observables, $z^{\alpha}_n = f(z^{\alpha}_{1n}; \pm_{1n}); : : : ; (z^{\alpha}_{nn}; \pm_{nn})g$, donde $z^{\alpha}_{kn} = h^{i\,1}(z_{kn})$ y $\pm_{kn} = 1$ si z_{kn} es positivo. Si $z_{kn} = 0$ entonces $\pm_{kn} = 0$ y z^{α}_{kn} es una observación censurada del tipo I y por la izquierda, con punto de censura igual a cero. De aquí la relación entre datos censurados y variables latentes.

Así, por la relación (4:1) ; la inferencia para el fenómeno de donde provienen los datos z_n se basa en la verosimilitud asociada a los datos z^{α}_n. De esta forma, si las variables aleatorias $fZ^{\alpha}_{1n}; : : : ; Z^{\alpha}_{nn}g$ son independientes e idénticamente distribuidas con función de densidad asociada g, la verosimilitud asociada a z^{α}_n la podemos expresar como

$$\prod_{k=1}^{Y} [g(z^{\alpha}_{kn})]^{\pm_{kn}} [G(0)]^{1\,i\,\pm_{kn}}; \tag{4.2}$$

donde G es la Función de Distribución respectiva a g.

Es importante tomar conciencia que aquí los datos observados del fenómeno o experimento bajo estudio no están censurados ni truncados, según las de…niciones dadas en la sección 1. Sin embargo, los datos correspondientes a las variables latentes son censurados de Tipo I (en la exposición de esta sección censurados de Tipo I y por la izquierda).

Ejempli…caré lo anterior con el modelo tobit. El modelo tobit, propuesto formalmente en [Tobin, 1958], es un modelo de regresión utilizado frecuentemente en Econometría.

EJEMPLO 4.1 El modelo tobit. Supongamos que los datos son de la forma $z_n = f(z_{1n}; w_{1n}); : : : ; (z_{nn}; w_{nn})g$, donde algunos de los valores z_{kn} son cero y el resto son positivos. Además, w_{kn} representa un vector de información conocida. Las variables aleatorias $Z_n = fZ_{1n}; : : : ; Z_{nn}g$ que modelan el comportamiento del fenómeno o experimento de donde provienen los datos z_n; heredan sus propiedades estadísticas de la variables aletorias latentes $Z_n^¤ = fZ_{1n}^¤; : : : ; Z_{nn}^¤g$.

El modelo tobit es un modelo de regresión lineal entre las variables latentes

y la información conocida $fw_{1n}; \cdots ; w_{nn}g$. Este es,

$$Z^{\alpha}_{kn} = {}^{-T} w_{kn} + \frac{3}{4}"_{kn}:$$

para $k = 1; \cdots ; n$. Además $^{-}$ es un vector de parámetros y $f"_{1n}; \cdots ; "_{nn}g$ es un conjunto de variables aleatorias independientes e idénticamente distribuidas con Función de Distribución, por lo general, $Normal\,(0; 1)$. $\frac{3}{4}^2$ representa la varianza de las variables aleatorias Z^{α}_{n}.

La relación entre las variables $fZ_{1n}; \cdots ; Z_{nn}g$ y las variables latentes es

$$Z_{kn} = \begin{cases} Z^{\alpha}_{kn} & si \quad Z^{\alpha}_{kn} > 0 \\ 0 & si \quad Z^{\alpha}_{kn} \quad 0: \end{cases}$$

Aquí, la función h es la identidad.

Siguiendo la verosimilitud expresada en la ecuación (4:2), la verosimilitud para el modelo tobit es

$$\prod_{k=1} \frac{1}{\frac{3}{4}} Á \left(\frac{Z^{\alpha}_{kn} - {}^{-T} w_{kn}}{\frac{3}{4}} \right)^{\pm_{kn}} \left(© \left(- \frac{{}^{-T} w_{kn}}{\frac{3}{4}} \right) \right)^{1 - \pm_{kn}};$$

donde Á es la función de densidad de una normal estándar y © es su Función

de Distribución respectiva. Además el conjuto de datos

$$z_n^{\alpha} = f(z_{1n}^{\alpha}; \pm_{1n}); : : : ; (z_{nn}^{\alpha}; \pm_{nn})g$$

es como se de…nio anteriormente.

Este ejemplo sirve para ilustrar, con un modelo especí…co, la relación entre un conjunto de datos que tienen algunos de sus valores concentrados en cero y los correspodientes datos censurados de las variables latentes asociadas.

Referencias básicas del modelo tobit la podemos encontrar en la sección 20.3 de [Greene, 1997] y el capítulo 6 de [Maddala, 1983].

En la sección siguiente expondré como se ha extendido la modelación de datos censurados con variables aleatorias independientes a datos espaciales censurados.

5 Datos espaciales

En fenómenos espaciales los instrumentos o procedimientos de medición son de precisión …nita, por lo que es de esperar tener frecuentemente datos censurados producidos por esta cualidad de los instrumentos o procedimientos referidos. Es conveniente mencionar nuevamente el trabajo de [Lambert et al., 1991] donde se expone la relación que existe entre los procedimientos de medición y los datos censurados, en el contexto de contaminación ambiental.

Los datos espaciales son de la forma $z_n = \{z(x_{1n}); \ldots; z(x_{nn})\}$; donde x_{kn} son las coordenadas espaciales donde fue registrado el dato $z(x_{kn})$.

Si el interes de estudiar el fenómeno de donde provienen los datos z_n es predecir alguna característica de la región de estudio, llamemosla a esta última A, entonces el modelo utilizado para describir el fenómeno es el campo aleatorio $Z = \{Z(x); x \in A\}$. Si lo que interesa estudiar es alguna característica localizada en una o varias de las coordenadas de observación $\{x_{1n}; \ldots; x_{nn}\}$, entonces el modelo es el conjunto de variables aleatorias $Z_n = \{Z(x_{1n}); \ldots; Z(x_{nn})\}$. Por último, si lo que interesa es predecir alguna característica en h puntos donde no se tienen observaciones, digamos

$\{x_1^0, \ldots, x_h^0\}$, entonces el modelo es

$$Z_{n+h} = \{Z(x_{1n}), \ldots, Z(x_{nn}), Z(x_1^0), \ldots, Z(x_h^0)\}.$$

Además, las variables aleatorias involucradas en el modelo que describe un fenómeno en la región A, son por lo general no independientes. Esto último contrasta con los modelos para datos censurados presentados en las anteriores secciones de este trabajo.

La censura producida por los instrumentos de medición es, por lo general, del Tipo I y por la izquierda o por la derecha o ambas (ver la sección 2). Esto es porque los mencionados instrumentos tienen sus límites de detección bien de…nidos. En la Introducción se han mencionado algunos ejemplos de este tipo de datos. En ocaciones los instrumentos de medición tienen resolución limitada, por lo que surge la censura por intervalos. Sin embargo este tipo de censura no se ha tratado en la literatura de datos espaciales censurados.

Los datos espaciales censurados se representan por

$$z_n = \{(z(x_{1n}), \pm_{1n}), \ldots, (z(x_{nn}), \pm_{nn})\},$$

con conjuntos de censura $\{B_{1n}, \ldots, B_{nn}\}$, donde las $\pm$s nos indican la censura

como antes.

La expresión algebraica de la función de densidad conjunta de las variables aleatorias Z_n, dada la información sobre censura $f\pm_{1n};:::;\pm_{nn}g$; es no trivial, aun en el caso de tener un modelo especí…co como el de la Función de Distribución normal multivariada.

Dada esta di…cultad, procederé a exponer como se tratan este tipo de datos según se muestra en los trabajos de [Militino y Ugarte, 1999] y [Stein, 1992].

Para exponer la metodología propuesta en [Militino y Ugarte, 1999] es conveniente recordar lo expuesto en la sección 3.3. Ahí, las variables aleatorias Z_n son independientes e idénticamente distribuidas y los datos z_n tienen censura del Tipo I. Siguiendo la misma sección, consideremos el caso de la censura por la izquierda, entonces la verosimilitud asociada a los datos z_n es

$$\prod_{k=1}^{Y} (g(z_{kn}))^{\pm_{kn}} (G(z_{kn}))^{1 i \pm_{kn}}; \tag{5.1}$$

donde g es la función de densidad asociada a Z_n y G es su respectiva Función de Distribución. Además, cuando $\pm_{kn} = 0$, en el caso de censura, los conjuntos de censura son de la forma $B_{kn} = (¡ 1; z_{kn})$:

Además supongamos que las variables aleatorias Z_n siguen un modelo de

regresión lineal de la forma

$$Z_{kn} = {}^{-T} w_{kn} + \frac{3}{4}"_{kn}; \tag{5.2}$$

donde $^{-}$ es un vector de dimensión q de coe...cientes y los ws son vectores de información conocida. También, las variables aleatorias $f"_{1n}; : : : "_{nn}g$ son independientes e idénticamente distribuidas con Función de Distribución asociada normal estándar y $\frac{3}{4}^2$ es la varianza de cada Z_{kn}: Así, la verosimilitud resultanted derivada de de la expresión (5:1) es

$$\prod_{k=1} \frac{1}{\frac{3}{4}} Á \left(\frac{z_{kn} - {}^{-T} w_{kn}}{\frac{3}{4}} \right)^{\P, \pm_{kn}} © \left(\frac{z_{kn} - {}^{-T} w_{kn}}{\frac{3}{4}} \right)^{\P, 1 - \pm_{kn}}; \tag{5.3}$$

donde Á representa la función de densidad de una normal estándar y © su Función de Distribución.

Bajo este escenario, una forma de obtener estimadores de los parámetros $(^{-}; \frac{3}{4})$ es máximizando la verosimilitud en la expresión (5:3) : La presentación del anterior modelo sirve para mencionar que una forma equivalente de estimación a la de máxima verosimilitud es la utilización del algoritmo EM, el cual se presenta a continuación.

El algoritmo EM de [Dempster et al., 1977] es un método para producir

una sucesión de estimaciones de parámetros que, bajo algunas condiciones de regularidad, converge a la estimación por máxima verosimilitud. La base del algoritmo es la utilización de la verosimilitud para datos completos, esto es sin censura; que no se tienen todos registrados.

Este método consiste de la repetición de dos pasos. En el paso E, llamado asi porque involucra valores esperados, cada dato censurado se reemplaza por su valor esperado condicionado a que pertenece al conjunto de censura; en el caso expuesto arriba, condicionado a que está censurado por la izquierda. En este paso se debe tener una estimación de los parámetros involucrados. En el paso M, llamado asi porque involucra la máximización de la verosimilitud, se estiman los parámetros involucrados con la verosimilitud completa; la verosimilitud con los datos no censurados y los datos censurados remplazados por los calculados en el paso E.

Especí…camente, con el modelo de la expresión (5:2), en la t ¡ ésima iteración del paso E y siendo $\hat{\beta}^{(t)}; \hat{\sigma}^{(t)}$ los valores de los parámetros de interes,

el k ¡ ésimo dato censurado se sustituye por

$$z_{kn}^{(t)} = E\left(Z_{kn} \mid Z_{kn} < z_{kn}\right) = \frac{E\left[Z_{kn} I_{(Z_{kn} < z_{kn})}\right]}{\Phi\left(\frac{z_{kn} - \beta^{(t)T} w_{kn}}{\sigma^{(t)}}\right)};$$

utilizando el modelo (5:2) se obtiene que

$$\begin{aligned} z_{kn}^{(t)} &= \frac{E\left[\beta^{(t)T} w_{kn} + \sigma^{(t)} \varepsilon_{kn} \; I_{(Z_{kn} < z_{kn})}\right]}{\Phi\left(\frac{z_{kn} - \beta^{(t)T} w_{kn}}{\sigma^{(t)}}\right)} \\ &= \frac{\beta^{(t)T} w_{kn} E\left[I_{(Z_{kn} < z_{kn})}\right] + \sigma^{(t)} E\left[\varepsilon_{kn} I_{\left(\varepsilon_{kn} < \left[z_{kn} - \beta^{(t)T} w_{kn}\right]/\sigma^{(t)}\right)}\right]}{\Phi\left(\frac{z_{kn} - \beta^{(t)T} w_{kn}}{\sigma^{(t)}}\right)} \\ &= \beta^{(t)T} w_{kn} + \sigma^{(t)} \frac{E\left[\varepsilon_{kn} I_{\left(\varepsilon_{kn} < \left[z_{kn} - \beta^{(t)T} w_{kn}\right]/\sigma^{(t)}\right)}\right]}{\Phi\left(\frac{z_{kn} - \beta^{(t)T} w_{kn}}{\sigma^{(t)}}\right)} \\ &= \beta^{(t)T} w_{kn} - \sigma^{(t)} \frac{\phi\left(\frac{z_{kn} - \beta^{(t)T} w_{kn}}{\sigma^{(t)}}\right)}{\Phi\left(\frac{z_{kn} - \beta^{(t)T} w_{kn}}{\sigma^{(t)}}\right)}: \end{aligned} \tag{5.4}$$

Este predictor también se puede encontrar en los trabajos de [Schmee y Hahn, 1979] y [Aitkin, 1981].

Es importante notar que $z_{kn}^{(t)}$ de la expresión (5:4) es formalmente

$$z_{kn}^{(t)} = E\left[Z_{kn} \mid f Z_{in} : \pm_{in} = 1g \setminus f Z_{jn} : \pm_{jn} = 0g \right] :$$

Pero por la independencia de las variables aleatorias Z_n, se tiene que

$$z_{kn}^{(t)} = E\left(Z_{kn} j Z_{kn} < z_{kn}\right); \tag{5.5}$$

dado que z_{kn} es un dato censurado.

Ahora, en el paso M se estiman los parámetros por máxima verosimilitud utilizando los datos completos. Esto es, los datos censurados han sido sustituidos por los valores esperados condicionales calculados en el paso E. Así, los nuevos parámetros estimados son

$$-^{(t+1)} = \left(W^T W \right)^{-1} W^T z_n^{(t)};$$

donde $z_n^{(t)}$ son los datos no censurados más los producidos en el paso E y W es la matriz cuyos renglones son los vectores de información w_{kn}^T: Además,

en este paso la estimación de σ^2 es

$$\hat{\sigma}^{2(t+1)} = \frac{\sum_{k=1}^{n} (z_{kn}^2)^{(t)} - 2z_{kn}^{(t)}\beta^{(t+1)}w_{kn} + \left(\beta^{(t+1)}w_{kn}\right)^2}{n - q};$$

donde $(z_{kn}^2)^{(t)}$ también es producido en el paso E y está dado por

$$\begin{aligned}\left(z_{kn}^2\right)^{(t)} &= E\left(Z_{kn}^2 | Z_{kn} < z_{kn}\right) \\ &= \left(\beta^{(t)T}w_{kn}\right)^2 + \hat{\sigma}^{2(t)} - \hat{\sigma}^{(t)}\beta^{(t)T}w_{kn} + z_{kn}\frac{\phi\left(\frac{z_{kn} - \beta^{(t)T}w_{kn}}{\hat{\sigma}^{(t)}}\right)}{\Phi\left(\frac{z_{kn} - \beta^{(t)T}w_{kn}}{\hat{\sigma}^{(t)}}\right)}.\end{aligned}$$

El calculo de $(z_{kn}^2)^{(t)}$ es muy similar al de $z_{kn}^{(t)}$ de la expresión (5:4):

El proceso de iteración término cuando los valores de las estimaciones de los parámetros en dos iteraciones sucesivas son muy parecidos.

Ahora regresemos al caso de datos espaciales con censura donde estos son de la forma $z_n = \{(z(x_{1n}); \pm_{1n}); \ldots; (z(x_{nn}); \pm_{nn})\}$; con conjuntos de censura $\{B_{1n}; \ldots; B_{nn}\}$.

En el trabajo de [Militino y Ugarte, 1999] utilizan un modelo lineal de la

forma

$$Z(x_{kn}) = \bar{}^{T} f_q(x_{kn}) + "(x_{kn}); \tag{5.6}$$

para $k = 1; : : : ; n$, donde $\bar{}$ es un vector de parámetros y f_q es un vector de dimensión q conocido de funciones de las coordenadas espaciales. Además, $f"(x_{1n}); : : : ; "(x_{nn})g$ es un conjunto de variables aleatorias con Función de Distribución conjunta normal multivariada con media cero y matriz de covarianzas $§_n$.

En el mismo trabajo de [Militino y Ugarte, 1999] se menciona que la correlación espacial, modelada por $§_n$, introduce serias di…cultades para el calculo de las esperanzas condicionales requeridas en el algoritmo EM.

Esta esperanza condicional mencionada en el referido artículo, que no la hacen explícita, es

$$E\begin{bmatrix} Z(x_{k_1 n}) \\ \vdots \\ Z(x_{k_r n}) \end{bmatrix} fZ(x_{in}) : \pm_{in} = 1g \setminus fZ(x_{jn}) : \pm_{jn} = 0g; \tag{5.7}$$

donde $fx_{k_1 n}; : : : ; x_{k_r n}g$ son las coordenadas donde se tienen los datos cen-

surados. Como no tenemos el supuesto de independencia para las variables Z_n no podemos obtener una expresión igual a (5:5). De aquí, posiblemente, la di…cultad mencionada para calcular las experanzas condicionales del paso E del algoritmo EM.

Como una alternativa, [Militino y Ugarte, 1999] propone transformar el modelo (5:6) para facilitar el uso del algoritmo EM. El modelo transformado que la mencionada referencia propone es expresado como

$$\tilde{Z}_n = \Gamma_n^{-} + \varepsilon_n; \tag{5.8}$$

donde $\varepsilon_n = (\varepsilon(x_{1n}); \ldots; \varepsilon(x_{nn}))^T$, $\tilde{Z}_n = \left[\tilde{Z}(x_{1n}); \ldots; \tilde{Z}(x_{nn})\right]^T$ y la matriz Γ_n es de la forma

$$\Gamma_n = \begin{pmatrix} 0 & & 1 \\ \gamma(x_{1n})^T & & \\ \vdots & & \\ \gamma(x_{nn})^T & & \end{pmatrix};$$

cuyas componentes se explican a continuación.

Cada componente de $\tilde{Z}_n$ es de la forma $\tilde{Z}(x_{jn}) = Z(x_{jn}) - Z_{n\,|\,j}^T K_{n\,|\,j}^{-1} k_j$ donde $Z_{n\,|\,j}$ es el vector de variables aleatorias Z_n después de haber re-

movido la j ésima componente. Además, $K_{n_{ij}}$ es la matriz de covarianzas del conjunto de variables aleatorias $Z_{n_{ij}}$ y k_j es el vector de covarianzas $Cov(Z(x_{jn}); Z_{n_{ij}})$. Por último, $\hat{f}(x_{jn}) = f(x_{jn}) - F^T_{n_{ij}} K^{-1}_{n_{ij}} k_j$. En [Militino y Ugarte, 1997] prueban que las variables aleatorias $\hat{\varepsilon}_n$, en el modelo transformado (5:8), son aproximadamente no correlacionadas. Por lo que aplican al modelo (5:8) el algoritmo EM para estimar el vector de parámetros θ, como se aplicó anteriormente al modelo (5:2).

Un enfoque diferente en el estudio de datos espaciales censurados se encuentra en el artículo de [Stein, 1992]. Ahí los datos espaciales censurados analizados son sobre la concentración de uranio en una perforación de exploración; además de datos simulados. El modelo utilizado para modelar las concentraciones de uranio fue (siguiendo la notación del presente trabajo)

$$Z(x) = \begin{cases} 0{:}18 \pounds \exp(Z^{\alpha}(x)) & \text{si} \quad Z^{\alpha}(x) > 0 \\ 0{:}1 & \text{si} \quad Z^{\alpha}(x) \quad 0{:} \end{cases} \tag{5.9}$$

En esta relación, $\{Z(x); x \in A\}$ es el campo aleatorio utilizado para modelar las concentraciones de uranio. Además, $\{Z^{\alpha}(x); x \in A\}$ es un campo aleatorio latente Gaussiano. Aquí A representa la perforación vertical donde se tomaron los datos de concentración de uranio.

Los datos de este problema son, efectivamente, datos censurados de Tipo I por la izquierda; con punto de censura igual a 0.1 (posiblemnte el límite inferior de detección del instrumento utilizado para las mediciones). Sin embargo, la inferencia fue hecha con base al campo aleatorio lantente y los datos censurados producidos por la relación (5:9).

Estos datos espaciales censurados se representan por

$$z_n = f(z(x_{1n}); \pm_{1n}); \ldots; (z(x_{nn}); \pm_{nn})g;$$

como antes, con conjuntos de censura $B_{jn} = (¡ 1; 0:1]$, $j = 1; \ldots; n$. Además, las $\pm$s nos indican la censura. Ahora, con base a la relación (5:9), los anteriores datos producen los datos no observados

$$z_n^¤ = f(z^¤(x_{1n}); \pm_{1n}); \ldots; (z^¤(x_{nn}); \pm_{nn})g;$$

donde $z^¤(x_{jn}) = \ln(z(x_{jn})=0:18)$ si $\pm_{jn} = 1$ y conjuntos de censura $B_{jn}^¤ = (¡ 1; 0]$, $j = 1; \ldots; n$.

Uno de los principales objetivos del artículo de [Stein, 1992] es la estimación de la Función de Distribución condicional de $Z(x)$, $x \; 2 \; A$, dado el

evento

$$E := \bigcap_{\pm_{kn}=1} fZ(x_{kn}) = z(x_{kn})g \bigcap_{\pm_{jn}=0} fZ(x_{jn}) \quad 0:1g:$$

Nuevamente bajo la relación (5:9), el anterior evento es igual a

$$E = \bigcap_{\pm_{kn}=1} \left[Z^{\alpha}(x_{kn}) = \ln\left(\frac{z(x_{kn})}{0:18}\right) \right] \bigcap_{\pm_{jn}=0} fZ^{\alpha}(x_{jn}) \quad 0g:$$

Explícitamente, la anterior función de distribución condicional evaluada en $t \in R$ es

$$\begin{aligned} & Pr(Z(x) \quad tj E) \\ = & Pr\left[Z^{\alpha}(x) \quad \ln\left(\frac{t}{0:18}\right) \right] - E \\ & \int_{(-1;0]^r} \left(\frac{\ln\left(\frac{t}{0:18}\right) - a - b^T v}{\cdot} \right) g(v|u)\, dv \\ = & \frac{}{\int_{(-1;0]^r} g(v|u)\, dv}: \end{aligned} \tag{5.10}$$

Aquí Φ es la función de distribución de una normal estándar. También, $u = \{z^{\alpha}(x_{jn}) : \pm_{jn} = 1\}$ es el vector de dimensión $n - r$ y conformado por los datos no censurados. Además, v es una variable de dimensión r, donde r es

el número de datos censurados. Con respecto a a y b, éstas son constantes que dependen de u. Aquí $¿^2$ es la varianza de la variable aleatoria $Z^¤(x)$. Por último, g es la función de densidad condicional normal multivariada del vector aleatorio conformado por $fZ^¤(x_{jn}) : ±_{jn} = 0g$.

El autor de [Stein, 1992] estima la Función de Distribución de la expresión (5:10) utilizando, primero, muestreo preferencial (ver la sección 3.3 de [Robert y Cassella, 1999]) y después un estimador de razón. Este procedimiento de estimación lo justi…ca por la di…cultad de evaluar las integrales involucradas en la expresión (5:10).

Una limitación de la metodología expuesta en el trabajo de [Stein, 1992] es que no puede aplicarse a conjuntos grandes de datos por la cantidad de cálculos computacionales requeridos; según lo expresa el autor en el mencionado artículo.

Más alla de lo expuesto en los trabajos de [Stein, 1992] y [Militino y Ugarte, 1999] no existe otro tipo de metodología sustancialmente diferente a la propuesta en dichos trabajos para modelar datos espaciales con censura.

En lo referente a variables latentes espaciales, exite la misma relación entre datos espaciales y variables latentes como la mostrada en la sección 4, con la salvedad que aquí se tiene una estructura de dependencia espacial (ver

el trabajo de [Stein, 1992]).

6 Comentarios …nales

En el presente trabajo se ha revisado la metodología para modelar y analizar datos con censura de fenómenos y experimentos; que involucran mediciones de vida útil o resistencia. Las variables utilizadas para modelar tales datos son variables aleatorias independientes e idénticamente distribuidas. Tal metodología ha recibido un gran número de contribuciones. En contraposición, la metodología para la modelación y análisis de datos espaciales con censura apenas ha dado sus primeros pasos.

Una forma de involucrarnos en el estudio de datos espaciales con censura es conocer primero los diferentes tipos de censura, como se presentaron en la sección 2. Debemos conocer también como se han modelado los diferentes tipos de datos censurados en el caso de modelación con variables aleatorias independientes, como lo expuesto en la sección 3. Esto último para explorar la posibilidad de extenderlo a datos que se modelan con variables que tienen una estructura de correlación espacial. Además es importante no confundir la modelación de datos censurados con la de datos truncados o la de variables latentes, el tratamiento númerico puede ser similar pero conceptualmente son cosas diferentes.

Antes de …nalizar con la exposición de este trabajo es oportuno presentar

una propuesta sobre modelación de datos espaciales censurados y una sobre la parte de estimación.

En la sección 5 se comento el trabajo de [Stein, 1992]. Ahí se utiliza censura, transformación de datos y variables latentes en el contexto de datos espaciales censurados. Si adicionalmente se añade un punto de concentración en los datos (ver la sección 4), el modelo correspondiente puede aplicarse a un número mayor de fenómenos espaciales. Más especí...camente, si los datos son censurados por la izquierda, con punto de censura $z_0 > 0$, y un punto de acumulación en 0, entonces un modelo posible a utilizar es

$$Z(x) = \begin{cases} h[Z^{\alpha}(x)] & \text{si} \quad Z^{\alpha}(x) > h^{i\,1}(z_0) \\ z_0 & \text{si} \quad 0 < Z^{\alpha}(x) \quad h^{i\,1}(z_0) \\ 0 & \text{si} \quad Z^{\alpha}(x) \quad 0; \end{cases}$$

Donde $fZ(x) ; x \; 2 \; Ag$ es el campo aleatorio utilizado para modelar las observaciones del fenómeno bajo estudio y A es la región de interes. Además $fZ^{\alpha}(x) ; x \; 2 \; Ag$ es un campo aleatorio latente (posiblemente Gaussiano) y h es una función positiva y monótona creciente sobre los realeas positivos. Nótese que este modelo puede utilizarse en datos sin referencia geográ...ca.

Un posible fenómeno donde se pueda aplicar el anterior modelo es la pre-

cipitación pluvial. Es común que se tengan registrados periodos de no precipitación (precipitación cero) y periodos con precipitación registrada (mayores que z_0). Adicionalmente es posible tener precipitaciones censuradas debidas al procedimiento de medición, bajas precipitaciones pero por debajo del límite de detección (mayores que cero pero menores o iguales a z_0).

Por último, si el objetivo es estimar los parámetros del modelo de datos espaciales censurados, se podría tratar de utilizar el algoritmo EM con tal modelo, sin necesidad de transformarlo como se sugiere en [Militino y Ugarte, 1999]. Esto ideando una forma operable de calcular la esperanza condicional en (5:7), fundamental para el paso E de tal algoritmo. O bien utilizar el enfoque de la sección 3 del trabajo de [Stein, 1992].

Referencias

[Aitkin, 1981] Aitkin, M. (1981). A Note on the Regression Analysis of Censored Data. Technometrics, 23, 161-163.

[Cohen, 1991] Cohen, A. C. (1991). Truncated and censored samples: theory and applications. Marcel Dekker.

[Dempster et al., 1977] Dempster, A. P., N. M. Laird y D. B. Rubin (1977). Maximun Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39, 1-38.

[Greene, 1997] Greene, W. H. (1997). Econometric analysis. Prentice-Hall.

[Kalb‡eisch y Prentice, 1980] Kalb‡eisch, J. D. y Prentice, R. L. (1980). The estatistical analysis of failure time data. John Wiley & Sons.

[Lambert et al., 1991] Lambert, D., B. Peterson y I. Terpenning (1991). Nondetects, detection limits, and the probability of detection. Journal of the American Statistical Association, 86, 266-277.

[Lawless, 1982] Lawless, J. F. (1982). Statistical models and methods for lifetime data. John Wiley Sons.

[Maddala, 1983] Maddala, G. S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge University Press.

[Militino y Ugarte, 1997] Militino, A. F. y Ugarte, M. D. (1997). A GM estimation of the location parameters in a spatial linear model. Communications in Statistics: Theory and Methods, 26, 1701-1725.

[Militino y Ugarte, 1999] Militino, A. F. y Ugarte, M. D. (1999). Analyzing censored spatial data. Mathematical Geology, 31, 551-561.

[Nelson, 1982] Nelson, W. (1982). Applied life data analysis. John Wiley & Sons.

[Robert y Cassella, 1999] Robert, C. P. y Cassella, G. (1999). Monte carlo statistical methods. Springer-Verlag New York.

[Schmee y Hahn, 1979] Schmee, J. y Hahn, G. J. (1979). A Simple Method for Regression Analysis with Censored Data (con discusión). Technometrics, 21, 417-432.

[Shumway et al., 1989] Shumway, R. H., A. S. Azari & P. Johnson (1989). Estimating mean concentrations under transformation for environmental data with detection limits. Technometrics, 31, 347-356.

[Stein, 1992] Stein, M. L. (1992). Prediction and inference for truncated spatial data. Journal of Computational & Graphical Statistics, 1, 91-110.

[Tobin, 1958] Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.