

Tarea 2 Estadística Espacial

Mario Becerra 124362

22/04/2015

Introducción

Las Montañas Azules son una cadena montañosa localizada en el noroeste de los Estados Unidos, que se extiende largamente por el este del estado de Oregón y el sudeste de Washington. En este trabajo se realiza un estudio sobre los incendios ocurridos en la zona con base en análisis estadístico-espacial, incluyendo variables como la vegetación y la elevación del terreno en la zona donde ocurre cada incendio. Se analiza el riesgo de incendio de distintos tipos de vegetación y se hace un análisis temporal del riesgo que existe en diferentes épocas del año.

Análisis exploratorio de datos

Antes de hacer cualquier tipo de modelo estocástico y antes de hacer cualquier tipo de prueba de hipótesis, es necesario analizar los datos que se tienen a la mano mediante un análisis exploratorio de datos. Esta parte es crucial en cualquier análisis estadístico pues en este primer paso se encuentra mucha información relevante sobre el fenómeno que se está trabajando.

Los datos que se usan en este proyecto corresponden a localidades de inicio de incendios forestales en la región de las Montañas Azules, en los estados de Oregon, Washington e Idaho. La información corresponde a incendios que comenzaron entre el 01 de abril de 1986 y el 31 de julio de 1993. Se tienen medidas las variables latitud, longitud, año, mes, día, tamaño, elevación, pendiente, orientación de la ladera donde ocurrió el incendio, días transcurridos desde el 1 de abril de 1986, y vegetación. Se pueden ver la presentación de los datos en la tabla 1.

% latex table generated in R 3.1.2 by xtable 1.7-4 package % Fri May 1 21:16:35 2015

	lon	lat	yr	mo	day	size	elev	slope	aspect	dia	veg9
1	764.49	816.11	86	4	14	0.10	1463	2	302.00	13	1
2	569.35	747.68	86	5	20	0.20	1310	3	26.00	49	5
3	542.17	700.37	86	5	26	0.10	1707	2	168.00	55	6
4	640.66	753.46	86	5	28	5.00	1400	8	228.00	57	5
5	510.35	646.56	86	5	29	11.00	1405	2	43.00	58	8
6	538.01	741.06	86	5	30	0.10	1404	2	252.00	59	1

Cuadro 1: Presentación de los datos disponibles

La dispersión de los incendios en el mapa se puede ver en la figura 1; donde se muestran los puntos de acuerdo a la elevación del terreno y el tamaño del incendio. Por el momento no se puede decir mucho sobre la relación que pueda existir entre las variables, excepto tal vez que al norte casi no hubo incendios y que en las zonas con menor elevación hay más incendios; sin embargo no son hipótesis que se puedan rechazar solamente viendo el mapa. La distribución de incendios pde acuerdo a la elevación se puede ver en la figura 2. El mayor número de incendios está entre 1150 m y 1920 m de altura.

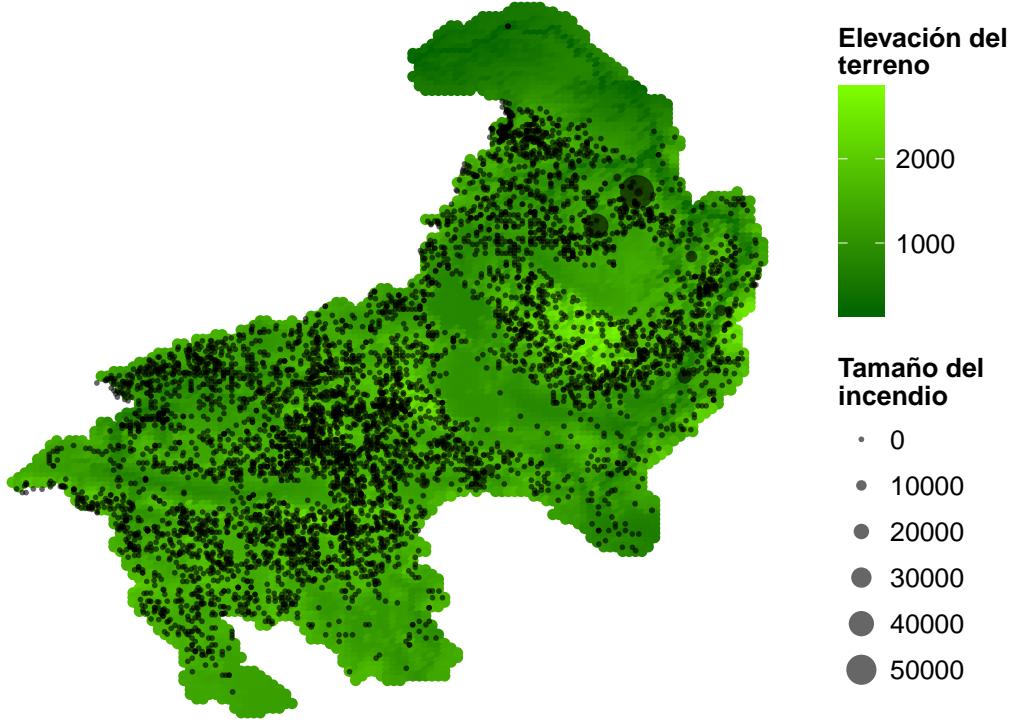


Figura 1: Dispersión de incendios en las Montañas Azules

Ahora, se analizará la elevación en el contexto temporal. En la figura 3 se aprecia la distribución del número de incendios de acuerdo a la elevación y al año. En cada año se mantiene la misma forma y la mayoría de los incendios ocurren a la misma altura en cada año, así que con solo ver la gráfica, se puede rechazar la hipótesis de dependencia temporal.

Si se divide a los datos por año y los se ven en el plano como en la figura 4, se ve que hubo mayor número de incendios en 1986 y menos en 1993, pero no parece haber un patrón muy evidente en la distribución espacial.

Un factor importante que puede afectar el número de incendios es el tipo de vegetación, pues algunos tipos de planta son más fáciles de encenderse y dispersarse que otros. Esto se puede ver en la figura 5, donde se aprecia que los tipos de vegetación 5,6 y 7 tienen mayor número de incendios.

Tal vez sea conveniente buscar alguna relación temporal en los incendios sin tomar en cuenta el factor espacial; para esto, agrupamos el número de incendios por año y por mes. En la figura 6 se puede ver, como ya se había mencionado, que hubo mayor número de incendios en

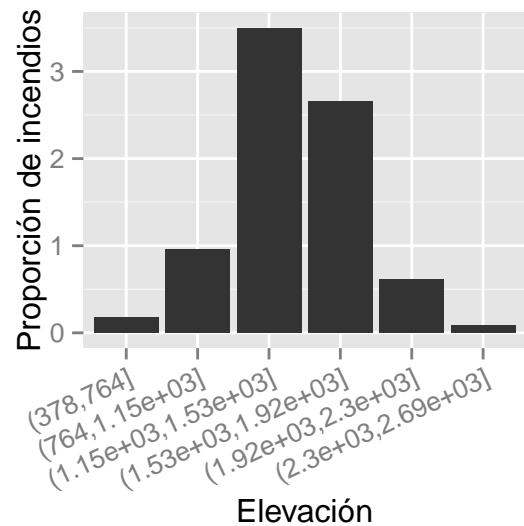


Figura 2: Distribución del número de incendios de acuerdo a la elevación

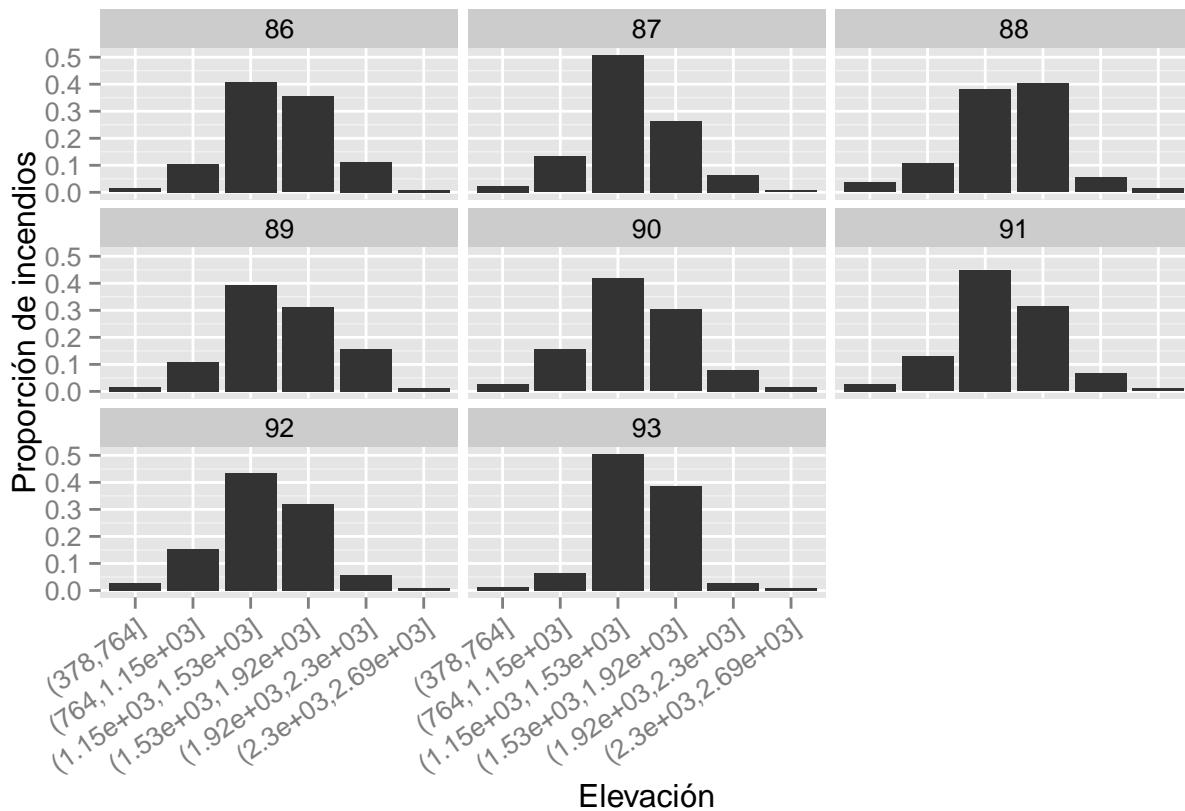


Figura 3: Distribución del número de incendios de acuerdo a la elevación por año

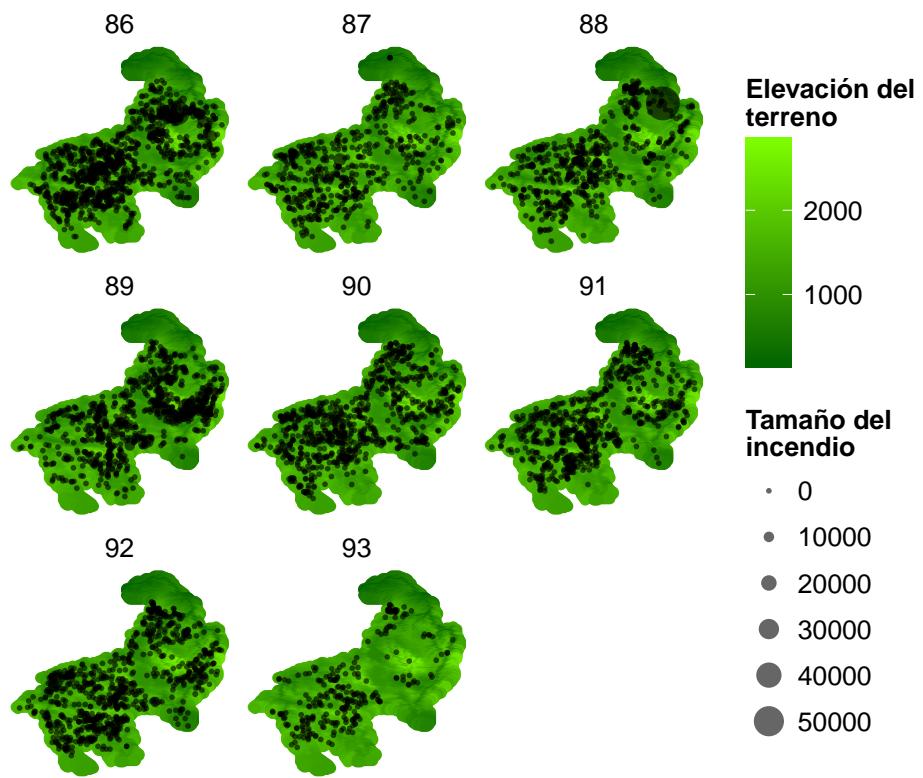


Figura 4: Dispersión de incendios en las Montañas Azules por año

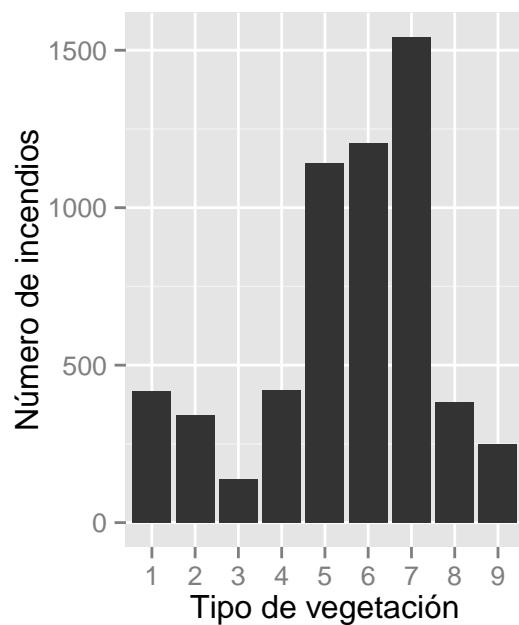


Figura 5: Distribución del número de incendios de acuerdo al tipo de vegetación

1986; y algo que es notable, pero no sorprendente, es que en los meses de verano hay mayor número de incendios.

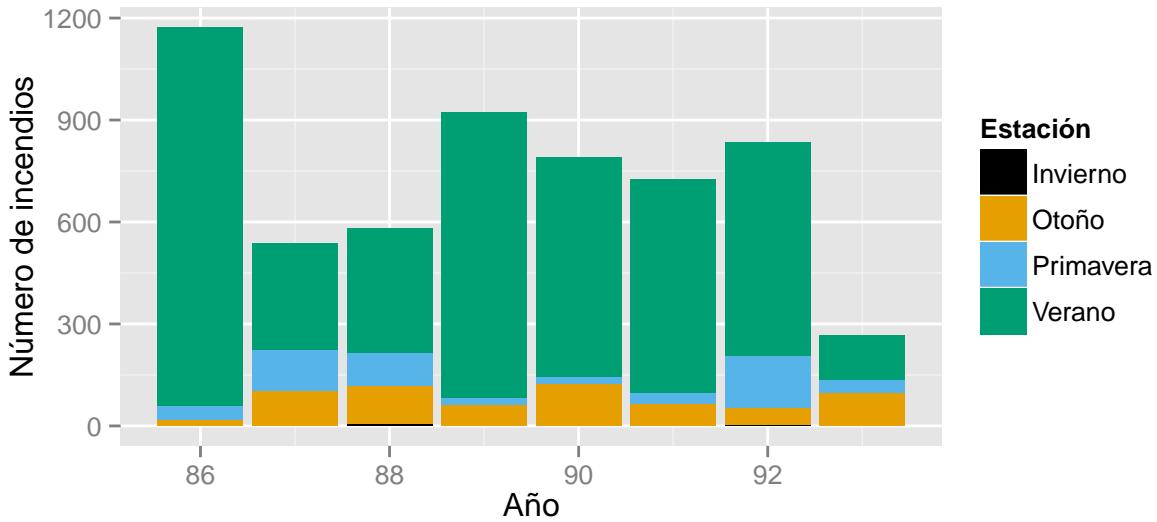


Figura 6: Número de incendios por año y por estación en las Montañas Azules

Se vio que hay mayor número de incendios en verano y que el mayor número de incendios provienen de vegetación tipo 5, 6 y 7; pero tal vez el riesgo de incendio cambie de acuerdo a la estación. La figura 7 muestra esta relación. Se descartaron los incendios sucedidos en invierno porque solo ocurrieron 10 en total; sin embargo, en las estaciones restantes se puede ver un patrón algo regular, aunque sí se puede notar que el número de incendios del tipo de vegetación 1 es mayor en primavera, y que el de tipo 4 es menor. Es difícil rechazar algún tipo de dependencia entre las variables de esta forma; para mayor confianza, se puede hacer una prueba *Ji-cuadrada*, pero como no es el interés en este estudio, no se lleva a cabo.

Metodología

Es de interés saber si los incendios se distribuyen al azar o si existe alguna dependencia espacial; puede ser que exista alguna tendencia a una configuración de conglomerados (*clusters*). Aunque ya se analizó un poco esto visualmente, en esta sección se irá más a fondo en el aspecto cuantitativo y con herramientas de estadística espacial se procederá a ajustar y probar modelos; en particular, modelos de procesos puntuales para datos espaciales.

Un punto de partida conveniente en el análisis de procesos puntuales espaciales es probar la hipótesis de **Aleatoriedad Espacial Completa** (AEC), la cual se puede definir la AEC como un proceso Poisson homogéneo (PPH) en \mathbb{R}^n , esto es, el número de puntos contenidos en cualquier región A , $N(A)$, sigue una distribución Poisson con media $\lambda|A|$; donde $|A|$ es el área de la región A y λ es el parámetro de intensidad del proceso y además los puntos en la región A se distribuyen de manera aleatoria e independiente con distribución uniforme en A . Esto significa que si esta hipótesis fuera cierta, entonces los eventos (incendios en este

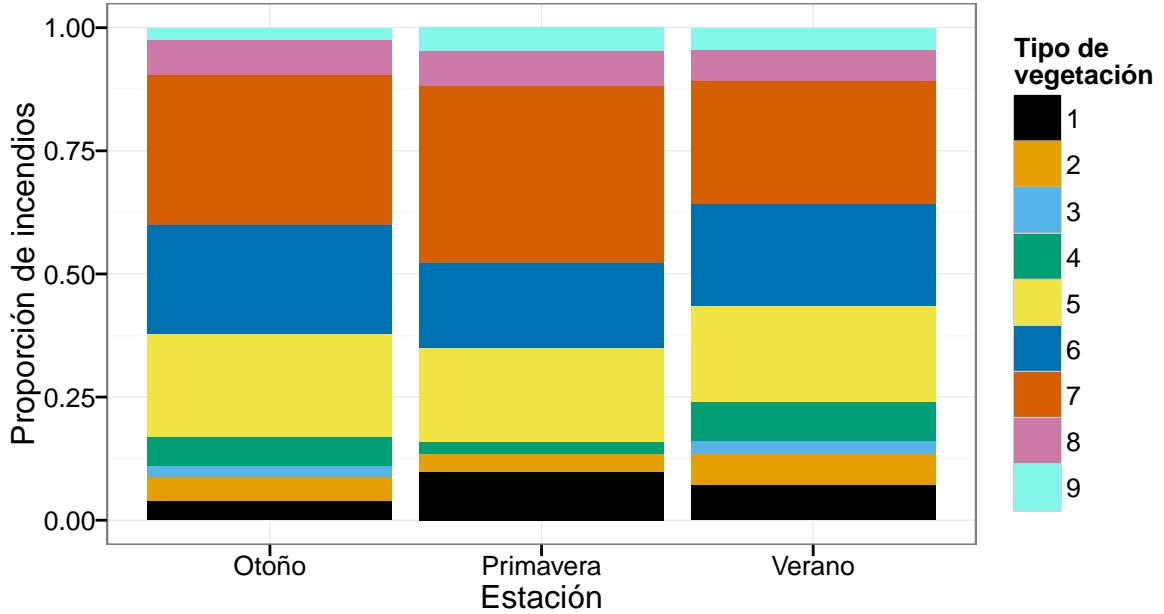


Figura 7: Proporción del número de incendios por estación en las Montañas Azules

caso) ocurren totalmente al azar, de forma constante en la región y que no hay interacción entre eventos.

Existen diversos estadísticos utilizados para probar la hipótesis de AEC; uno particular está basado en conteos de eventos en cuadrantes (áreas bien definidas, usualmente rectangulares) en la región de interés A . Otro estadístico está basado en distancias entre eventos, en específico, el vecino más cercano, ya sea desde un punto x del patrón observado, o desde un punto arbitrario. Y otro más es la K de Ripley, la cual se define como $K(h) = \frac{1}{\lambda} \mathbb{E}[\# \text{ extra de eventos dentro de una distancia } h \text{ a un evento arbitrario}]$.

Para el primer caso (conteos), si suponemos una partición del espacio de interés en m cuadrantes y en cada uno hay n_1, n_2, \dots, n_m eventos, un estadístico natural es el **índice de dispersión** definido como $I = \sum_{i=1}^m \frac{(n_i - \bar{n})^2}{(m-1)\bar{n}}$, que bajo AEC debe tomar valor igual a 1.

Otro estadístico que se usa es $I' = \frac{(m-1) \sum_{i=1}^m (c_i - \bar{c})^2}{\bar{c}} = (m-1)I$. Bajo AEC $I' \sim \chi^2_{(m-1)}$, por lo que se rechaza la hipótesis de AEC al nivel de significancia α si $I' > \chi^2_{(m-1)(1-\alpha)}$.

Para el caso del método basado en distancias se define la variable aleatoria D como la distancia de un evento arbitrario al evento más cercano, entonces, bajo AEC,

$$\mathbb{P}(D > d) = 1 - e^{-\lambda\pi d^2}.$$

Entonces la media y la varianza de D son $\mathbb{E}[D] = \frac{1}{2\sqrt{\lambda}}$ y $Var[D] = \frac{4-\pi}{4\lambda\pi}$. Por esto, si se defina \bar{D} como la media muestral de las distancias, asumiendo n v.a.i.i.d., se tiene que

$\mathbb{E}[\bar{D}] = \frac{1}{2\sqrt{\lambda}}$ y $Var[\bar{D}] = \frac{4-\pi}{4n\lambda\pi}$; por lo que centrando

$$Z = \frac{\bar{D} - 1/(2\sqrt{\lambda})}{\sqrt{(4-\pi)/(4n\lambda)}} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

Así, si n es grande, el IC para AEC tendrá la forma $\bar{D} \pm Z_{1-\alpha/2}\sqrt{(4-\pi)(4n\lambda)}$.

En el caso de la K de Ripley, si hubiera AEC entonces $K(h) = \pi h^2$, pues el número de puntos dentro de un radio h debe ser proporcional al área del círculo de radio h . Si los datos estuvieran en conglomerados, uno esperaría que $K(h) > \pi h^2$, mientras que si hubiera algún tipo de repulsión se esperaría que $K(h) < \pi h^2$. La versión muestral de la K de Ripley es

$$\hat{K}(h) = \frac{|A|}{n^2} \sum_{i=1}^n \sum_{i \neq j} \frac{I_h(d_{ij})}{w_{ij}}$$

donde m es el número de eventos en A , w_{ij} es la proporción del círculo con centro en i y que pasa por j que está dentro de A , d_{ij} es la distancia entre los puntos i y j , I es la función indicadora para la distancia d_{ij} .

Muchas veces se usa la función $L(h) = \sqrt{\frac{K(h)}{\pi}} - h$, pues la varianza de L es aproximadamente constante bajo AEC. En la práctica se grafica $t - \hat{L}(t)$ contra t , la cual, en el caso de AEC, deberá ser aproximadamente una línea horizontal en el cero.

Si se rechaza la hipótesis de AEC, se deben considerar procesos no homogéneos. La extensión más simple es el Proceso Poisson no homogéneo (PPNH), el cual cumple los mismos principios de el PPH, excepto que la función de intensidad depende del sitio, $\lambda(x)$. Entonces, para un área $B \subset A$, se tiene que $\mathbb{E}[N(B)] = \int_B \lambda(u)du$ y $\mathbb{P}(N(b) = n) = \frac{[\int_B \lambda(u)du]^n \exp^{\int_B \lambda(u)du}}{n!}$. A este modelo se le pueden agregar más covariables referentes al sitio; por ejemplo, la elevación y la humedad del sitio.

Resultados y discusión

Para el caso de estudio de este trabajo se llevaron a cabo dos casos distintos para la primera prueba de conteos, uno con 50 particiones y otro con 100 particiones. Notar que estas particiones se hacen sobre un cuadrado, pero el área de las Montañas Azules es irregular, por lo que al final no se tienen 2500 y 10000 respectivamente, sino menos, en particular, se tienen 1235 y 4684. Para el caso con 50 particiones se tuvo que $I_{50} = 5.85$ y para 100 se tiene $I_{100} = 2.49$. Como en ambos casos el índice es mayor a uno, podemos rechazar AEC bajo este esquema.

Para la segunda prueba, con 50 $I'_{50} = 7217.54$ y con 100 $I'_{100} = 1.17 \times 10^4$, y para una $\alpha = 0.01$, se tiene que $\chi^2_{(m-1)(1-\alpha)} = 1153.44$, por lo que con esta prueba también se rechaza AEC.

También se llevó a cabo la prueba de distancias y se llegó al IC $[1.31, 1.36]$ con una $\bar{D} = 1.75$. Claramente $\bar{D} \notin [1.31, 1.36]$, por lo que también con esta prueba se rechaza AEC.

Las pruebas anteriores se hicieron para los datos completos, es decir, incluyen a todos los años, pero si se separaran por año podríamos ver si existe algún patrón temporal en los incendios. Para los casos de cada año ya no se hicieron dos casos, sino que se usaron solo 50 particiones.

Para cada año se tienen los siguientes índices de dispersión: $I_{89} = 2.64$, $I_{90} = 1.87$, $I_{91} = 2.35$, $I_{92} = 2.16$, $I_{93} = 1.8$. Todos son mayores que 1, por lo que se rechaza la hipótesis de AEC en cada año.

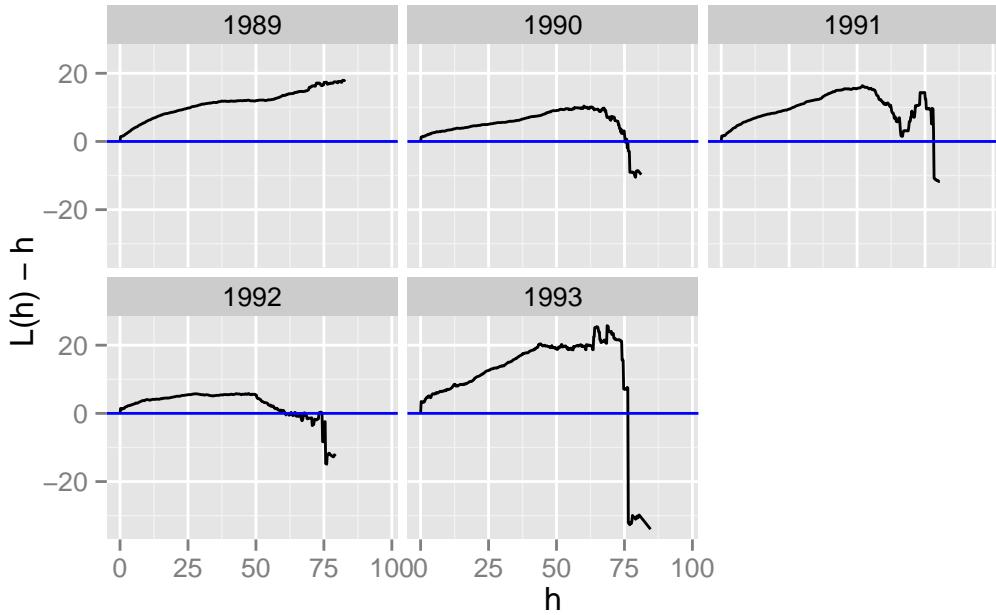


Figura 8: La función $L(h)$, transformación de la K de Ripley para cada año de estudio. En todas está lejos de ser una línea horizontal.

También se calculó la función $\hat{L}(h)$ para cada año y se graficaron los resultados, los cuales se pueden ver en la figura 8. En este estudio también se rechaza AEC pues las gráficas no representan una línea recta horizontal. Además, en esa misma figura se observa que $\hat{K}(h) > \pi h^2$ para la mayoría de los valores de h , por lo que se puede decir que los incendios tienden a ser en conglomerados. Esto tiene sentido pues la vegetación y las condiciones propicias para incendio usualmente se comportan de esta manera.

Después de diversas pruebas, se concluye que en este caso no existe AEC. El siguiente paso es usar un modelo que se ajuste a los datos. Como se mencionó antes, el proceso Poisson no homogéneo es la alternativa más sencilla al proceso de Poisson homogéneo. Se ajustaron tres modelos distintos para distintas estaciones del año, esto es, un modelo para verano, uno para primavera y uno último para otoño. Se descartó invierno por el muy bajo número de incendios que se tienen.

% latex table generated in R 3.1.2 by xtable 1.7-4 package % Fri May 1 21:17:22 2015

	Variable	Primavera	Verano	Otoño
1	(Intercept)	-6.0455	-4.5907	-6.4547
2	elev	0.0008	0.0011	0.0005
3	veg2	0.3144	0.9371	1.5863
4	veg3	-2.1611	0.3994	1.4133
5	veg4	0.0490	0.9965	2.1328
6	veg5	1.0946	1.1414	2.2086
7	veg6	0.9322	1.1416	2.2091
8	veg7	1.0725	0.8351	1.9053
9	veg8	0.8040	0.8122	1.8772
10	veg9	-0.4654	-0.2850	-0.1966
11	slope	-0.0837	-0.0517	-0.0886

Cuadro 2: Coeficientes de los modelos ajustados

En la tabla 2 se muestran los coeficientes estimados para cada uno de los modelos, y en la figura 9 se muestra una gráfica de estos mismos coeficientes. Se puede ver que para los tres modelos el coeficiente de la elevación está muy cerca de cero; por lo que se puede decir que esta no afecta mucho al número de incendios. Se puede observar que para los distintos tipos de vegetación los coeficientes varían entre los modelos, y que además los que mayores coeficientes tienen en los tres modelos son los tipos 4, 5 y 6, aunque en el análisis exploratorio de datos se vio que el tipo 7 tenía más número de incendios. Es natural ver que los coeficientes de vegetación en primavera sean menores que en verano, y que los de otoño sean mayores a los de verano, pues la vegetación es más seca en otoño que en verano, y en verano que en primavera, por lo que las condiciones son más propicias para que haya incendios.

Conclusiones

En este estudio se profundizó mucho en el análisis exploratorio de datos pues es una herramienta esencial para el entendimiento del fenómeno, y para que a la hora de ajustar modelos se pueda tener conocimiento sobre si estos hacen sentido y son congruentes con los datos. En este caso así fue, los modelos que se ajustaron eran consistentes con lo que se vio en el análisis exploratorio. Uno de los primeros resultados que se nota en los modelos es la diferencia que existe en el riego de incendio de acuerdo a la época del año, aunque esto no debe sorprender pues es natural pensar que en época de calor hay más facilidad de incendio.

Otro resultado que se vio fue que los incendios tienden a estar en conglomerados o *clusters*, se cree (pues no se tiene apoyo de conocimiento experto en este tema) que esto es porque la vegetación propicia para incendio tiende a estar cerca en forma de conglomerados, aunque este resultado no es concluyente, aunque se podría continuar con este estudio para probar esta hipótesis.

También se vio en los modelos que los tipos de vegetación 4, 5 y 6 son más propensos a incendios, y esta propensión es mayor en otoño y menor en primavera. En los modelos no se notó que existiera mucha diferencia en el riesgo de incendios de acuerdo a la elevación, sin

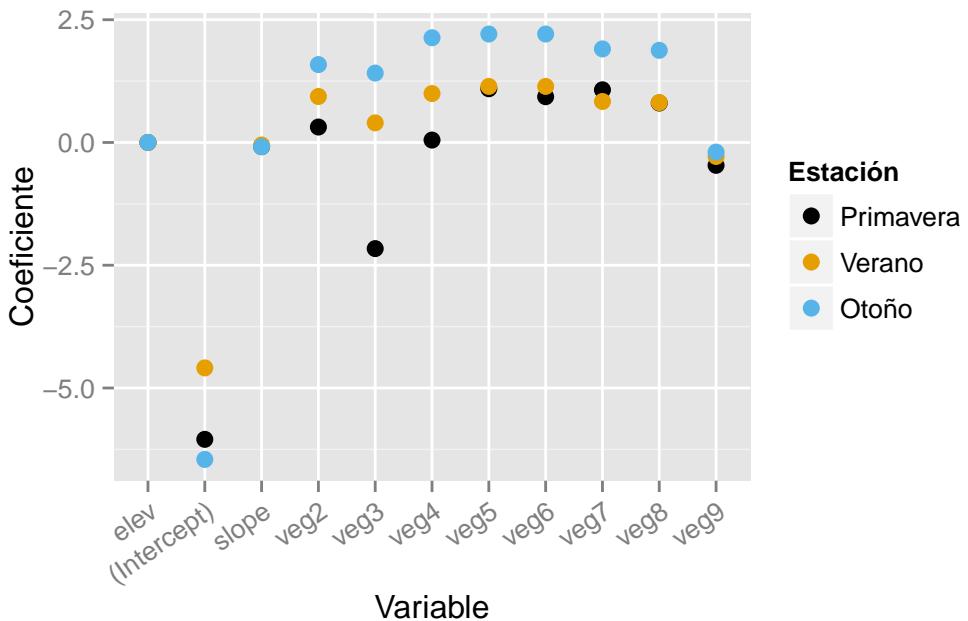


Figura 9: Coeficientes estimados de las variables en cada modelo.

embargo, en el análisis exploratorio de datos se vio que el mayor número de incendios ocurre entre 1150 m y 1920 m de altura.

Bibliografía

CRESSIE, N., en *Statistics for spatial data*, John Wiley & Sons, Inc., 1993.

DIXON, P. M., en *Ripley's k function*, John Wiley & Sons, Inc., 2002.

SUDIPTO BANERJEE, A. E. G., Bradley P. Carlin, en *Hierarchical modeling and analysis for spatial data*, Chapman; Hall/CRC, 2004.