Interpolación con Splines Naturales

Alfredo Méndez, Antonio Ramírez,Ariana López,
Christian Cuéllar, Eduardo Martínez
ITAM
23 de Septiembre de 2015

1 Motivación

¿Qué es un dato faltante?
Un dato faltante se presenta cuando no existe un valor almacenado para una observación en una variable dada. Es un fenómeno común en grandes bases de datos y en bases que reportan los resultados de encuestas.

¿Por qué hay datos faltantes?
- Errores del capturista.
- En una encuesta, el entrevistado decidió no responder a una o varias preguntas.
- No puede clasificarse una situación en variables categóricas.

¿Por qué es un problema?
En general, los datos faltantes dificultan la habilidad para procesar datos y encontrar resultados concluyentes. Entre los problemas asociados a los datos faltantes se encuentran:

Sesgo de estimadores
Si los valores faltantes no son resultado de un proceso aleatorio, los estimadores podrían estar sesgados.

Representatividad
Los datos faltantes reducen la representatividad de una muestra y, por lo tanto, podrían distorsionar las inferencias sobre la población.

2 Naturaleza de los datos faltantes

Datos faltantes completamente aleatorios (MCAR)
Este caso ocurre si los eventos que hicieron que un dato sea faltante son independientes tanto de las variables observables como de los parámetros de interés, y ocurren de manera completamente aleatoria. Si suponemos que hay algunos

valores faltantes en Y y que X es un vector de variables observadas, entonces se dice que los datos faltantes son MCAR si:

$Pr(Y \text{ tenga valores faltantes } |X, Y) = Pr(Y \text{ tenga valores faltantes})$

Por lo tanto, la probabilidad de que Y tenga un valor faltante no está relacionada con Y o con otras variables en X. En este caso, no hay sesgo y los datos siguen siendo representativos de la población. Sin embargo, los datos faltantes rara vez se comportan de esta forma.

Datos faltantes aleatorios (MAR)
Ocurre cuando los datos faltantes no se generan de manera aleatoria, pero la falta de estos datos puede ser totalmente explicada por variables en las que hay información completa. Por lo tanto, el patrón de los datos faltantes podría ser definido a partir de otras variables.

Los datos faltantes en Y son MAR si:

$Pr(Y \text{ tenga valores faltantes } |X, Y) = Pr(Y \text{ tenga valores faltantes } |X)$

Por lo tanto, la probabilidad de que Y tenga un valor faltante no depende de Y, tras controlar por variables observadas. No es posible verificar estadísticamente el supuesto de datos faltantes tipo MAR, por lo que éste debe ser razonable.

Datos faltantes no aleatorios (NMAR)
En este caso, el valor de la variable que es faltante se relaciona con la razón por la que es faltante. En otras palabras, los valores faltantes dependen de otros valores faltantes, por lo que no se puede realizar imputación de datos existentes. Un ejemplo sería que los hombres no pudieran completar una encuesta de depresión debido a su nivel de depresión.

En ocasiones es necesario tratar o completar los valores faltantes de una base de datos. Los métodos para realizar estas tareas incluyen la remoción de tuplas con valores desconocidos; la asignación de un valor relacionado con la distribución de los datos conocidos como puede ser la media, la moda, la mediana, o los valores máximos o mínimos; estimaciones de valores ajustados por técnicas de máxima verosimilitud; entre otras que se detallarán más adelante. Rellenar estos huecos en las bases se relaciona con el concepto de interpolación que se describe en el siguiente apartado.

3 Interpolación y otros conceptos relacionados

Interpolación vs. Extrapolación

La interpolación es un método para construir nuevos datos dentro del rango de un conjunto de datos conocidos (también llamados nodos o puntos fijos). En

este sentido, la interpolación construye una función f(x) que se ajusta a una serie de puntos conocidos (xk,yk). Dada una secuencia de n+1 distintos números xk con correspondientes números yk, se busca una función tal que
$f(xk) = yk, \quad k=0,1,\ldots,n$
Cada par (xk,yk) es un punto observado y a la f se le conoce como el interpolante para los puntos. A diferencia de la interpolación, la extrapolación se refiere al proceso mediante el cual se estima el valor de una variable con base en la relación que tiene con otra fuera del rango original observado.

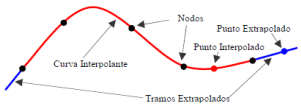


Figure 1: Curva Interpolante

Colocación vs. Aproximación

La colocación se refiere a la estimación del comportamiento de una función que relaciona las variables y necesariamente pasa por todos los puntos observados o nodos. La técnica de interpolación con splines naturales, que se detalla más adelante, representa un ejemplo de colocación de datos. Por su parte, en el proceso de aproximación se estima una función que relaciona las variables y la cual puede o no pasar por los datos observados. La línea de regresión en una estimación de Mínimos Cuadrados Ordinarios representa un ejemplo de aproximación.

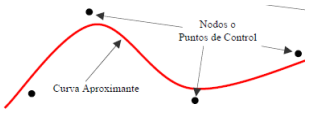


Figure 2: Curva aproximante

Métodos de interpolación:

Hay muchos métodos de interpolación distintos. Las cuestiones relevantes al momento de elegir un método apropiado son su exactitud, qué tan suave (diferenciable) es el interpolante, cuántos puntos se necesitan para interpolar, así como su simplicidad.

Con respecto a los puntos necesarios para interpolar, las técnicas de interpolación se dividen en dos categorías de acuerdo a la cantidad de nodos que utilizan: globales y locales. Los métodos globales consideran todos los nodos;

en cambio para los métodos locales sólo se consideran los nodos cercanos y a los más lejanos se les asigna peso nulo. En general, se utilizan los métodos locales pues son algorítmicamente más veloces, sobre todo cuando se dispone de un método rápido para encontrar el conjunto de nodos cercanos. A continuación se presentan algunos ejemplos de métodos de interpolación:

Interpolación por el método del vecino más cercano.

Es uno de los métodos más simples. Consiste en localizar el valor del dato más cercano y asignar el mismo valor. Este método pocas veces es utilizado en problemas unidimensionales, pero es una opción atractiva conforme aumenta el número de dimensiones por su velocidad y simplicidad.

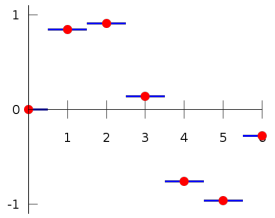


Figure 3: Una dimensión

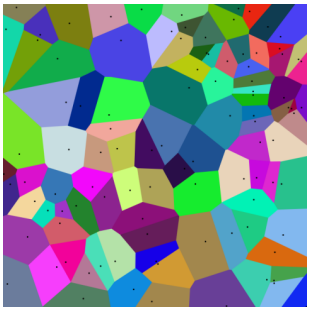


Figure 4: Dos dimensiones

Nota: los puntos representan los valores conocidos.

Interpolación lineal

Este método construye líneas rectas entre dos nodos cuyos valores son conocidos. Si dos puntos conocidos se presentan con las coordenadas (x_0, y_0) y

(x_1, y_1), entonces a un valor x en el intervalo (x_0, x_1) se le asocia un valor y definido por la ecuación.
$\frac{\boldsymbol{y}-\boldsymbol{y}_0}{\boldsymbol{x}-\boldsymbol{x}_0} = \frac{\boldsymbol{y}_1-\boldsymbol{y}_0}{\boldsymbol{x}_1-\boldsymbol{x}_0}$
Por lo que,
$\boldsymbol{y} = \boldsymbol{y} - (\boldsymbol{y}_1 - \boldsymbol{y}_0)\frac{\boldsymbol{x}-\boldsymbol{x}_0}{\boldsymbol{x}_1-\boldsymbol{x}_0}$

La ecuación anterior es la fórmula de la interpolación lineal en el intervalo (x_0, x_1). La interpolación lineal tiene la ventaja de ser sencilla y rápida, pero no es muy precisa para aproximar funciones. Otra desventaja es que el interpolante no es diferenciable en el punto x_k.

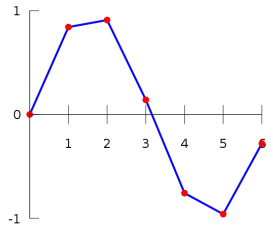


Figure 5: Interpolación lineal

- Interpolación polinomial

Es la generalización de la interpolación lineal y consiste en buscar un polinomio que pase exactamente por los puntos conocidos.

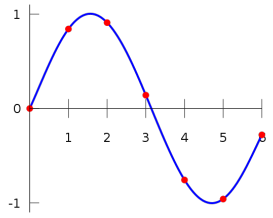


Figure 6: Interpolación Polinomial

En general, si se tienen n+1 puntos, entonces existe un polinomio de grado n que pasa por todos ellos. Sin embargo, conforme aumenta n, también se incrementa el número de ecuaciones. Además, el comportamiento de la curva interpolante

puede llegar a ser inestable, en el sentido que los valores interpolados podrían estar muy alejados de los valores observados. Esto es consecuencia del ajuste que debe hacer la curva interpolante para que pueda colocar los puntos conocidos, lo que provoca importantes oscilaciones en el intervalo de interpolación. Estos problemas son resueltos a través de la interpolación vía splines naturales, como se verá con mayor detalle más adelante.

4 Promedios de Ventanas Móviles

Una ventana es un conjunto delimitado de 'k' datos, donde el tamaño de la ventana es determinada por el usuario con base en la información que esté analizando, por ejemplo, precios diarios de una acción en un año, coordenadas en un área, datos transmitidos en cierta unidad de tiempo o el registro de las partículas contaminantes en el aire de cierta ciudad.

Dentro de una ventana móvil se pueden realizar diversos cálculos usando los valores conocidos en esta. En el contexto de llenar datos faltantes en bases de datos, algunos de los métodos utilizados para este fin son los promedios móviles, regresiones lineales, máximos y mínimos, valor precedente, valor subsecuente, uso de constantes, entre otros. (ver sección 3 para mayor referencia)

El método que explicaremos en esta sección, es el *promedio móvil*.

El término móvil indica que conforme se tenga disponible una nueva observación de los datos, se reemplaza la observación más antigua en el cálculo y se genera un nuevo promedio.

Un promedio móvil con un ancho de ventana 'k', promediará cada conjunto de k datos pasando por todos los datos disponibles. Los promedios obtenidos durante el movimiento de la ventana pueden ser utilizados para llenar huecos en la serie de datos originales.

No existe una regla específica que indique cómo seleccionar el tamaño de la ventana, sin embargo se recomienda que si la variable a promediar no tiene variaciones considerables entonces se utilice una 'k' grande, por el contrario, si la variable muestra patrones cambiantes es recomendable usar una 'k' pequeña.

En la figura 7, se presenta un ejemplo del movimiento del promedio a través de la serie de datos usando una ventana de 5 datos (k=5), el punto rojo y verde señalan el ultimo valor de la ventana que se está promediando.

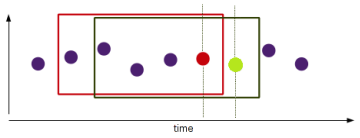


Figure 7: Ventana móvil

Una desventaja de los promedios móviles es que pierden la primera y última información de la serie de datos ya que introducen cierto retraso en su calculación mientras se alcanza el ancho de la primera ventana. La serie de promedios creados se considera una suavización de los datos originales.

Existen diversas formas para calcular promedios en una ventana móvil, aquí algunos de ellos:

Promedio móvil simple previo

Es una suma igualmente ponderada de k datos anteriores, donde cada peso es igual a 1/k para cada dato. La ventana de ancho k, se establece sobre el valor disponible más reciente de la serie.

$$PM_t = \frac{1}{k} * (X_t + X_{t-1} + ... + X_{t-k+1})$$

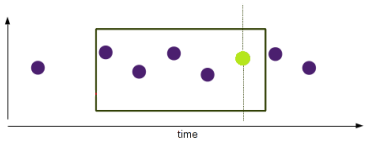


Figure 8: Promedio previo

Ventana Previa (k=5)

t-4	t-3	t-2	t-1	t

Promedio móvil simple centrado

En este tipo de promedios el valor del promedio móvil en el tiempo t, se calcula centralizando la ventana alrededor del tiempo t y promediando alrededor de los k valores de la ventana.

$$PMC_t = \frac{1}{k} * (X_{t-k+1} + ... + X_{t-1} + X_t + X_{t+1} + ... + X_{t+k-1})$$

Por ejemplo en una ventana de tamaño k=5, el promedio móvil en el tiempo t=3 significa promediar los datos en los tiempo 1,2,3,4,5; el promedio móvil en el tiempo t=4 significa promediar los datos en los tiempos 2,3,4,5,6, y así sucesivamente.

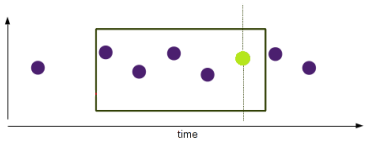


Figure 9: Promedio centrado

Ventana Centrada (k=5)

t-2	t-1	t	t+1	t+2

Promedio móvil ponderado
Es un promedio que contiene factores multiplicativos para dar diferentes pesos a los datos dentro de una ventana de modo que algunos datos tengan mayor relevancia que otros.

$$PMP_t = (W_1 * X_t) + (W_2 * X_{t-1}) + ... + (W_k * X_{t-k+1})$$

Dónde:

$\sum_{i=1}^{k} W_i = 1, W_i > 0$

Los pesos pueden ser elegidos para darle mayor importancia a los datos centrales, iniciales o finales de nuestra ventana.

En resumen, las consideraciones al trabajar con promedios móviles son:
- Suavizan la función de los datos
- Introducen cierto retraso en su calculación
- Pierde los datos al principio y al final de las series
- Sensible a los valores atípicos
- El error cuadrático es mayor en este método que con splines naturales
- No existe una regla específica que indique cómo seleccionar el tamaño de la ventana

Los promedios de ventanas móviles para completar datos faltantes en las bases de datos son muy usados debido a su facilidad de cálculo, sin embargo, es importante decir que no son la mejor herramienta para esta tarea, en la sección de splines de este documento se podrá observar que una mejor opción para interpolación de datos se encuentra en los splines naturales.

5 Definición de Splines

Un Spline es una función polinomial definida por “pedazos“. Es decir es la suma de varios polinomios, cada uno de ellos definido en un intervalo del dominio de

la función. La suma de todas estas funciones polinomiales es llamada un Spline. Mas formalmente definimos un spline como:

$S : [a, b] \to \mathbb{R}, \qquad [a, b] \subset \mathbb{R}$

en donde el intervalo [a,b] se encuentra particionado en subintervalos:

$[t_{i-1}, t_i] \subset [a, b]$

$a = t_0 < t_1 < ... < t_n = b$

y con la caracterísica de que la función S restringida a cada intervalo $[t_{i-1},t_i]$ es un polinomio

$S_i : [ti - 1, ti] \to \mathbb{R}$

Definiremos como **nudos** a los n puntos en los que encontramos delimitados los dominios de los polinomios S_i . Los llamaremos nudos ya que representan los puntos de unión de los distintos polinomios S_i.

La interpretación que podemos dar a esta definición es la de una función (llamémosla g) que "adivina" los valores de una función desconocida en los intervalos en los que no conocemos los valores de ésta, Conocemos los valores de la función g en los nudos únicamente y queremos decir algo de los valores de g en los espacios entre los nudos.

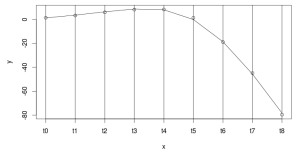


Figure 10: Ejemplo de un Spline

6 Tipos de Splines

P**olinomios de LaGrange**.

Si seguimos la definición de splines podemos fácilmente llegar a la conclusión de que los polinimios de grado n, es decir, las funciones de la forma:

$p(x) = a_0 + a_1x + ... + a_{n-1}x^{n-1} \in \Pi_{\prec n}$

Son casos particulares de splines. Esto nos lleva a la primera aproximación lógica teniendo el problema de interpolar puntos en un intervalo [a,b] que sería definir un sólo polinomio que coloque los puntos que tenemos dentro del intervalo. Esta aproximación es conocida como la aproximación de LaGrange y consiste en definir un sólo polinomio con las siguientes características:

$p(x) = \sum_{i=1}^{n} g(t_i)l_i \in \Pi_{\prec n}$

un polinomio de grado n. En donde

$l_i = \Pi_{j=1}^{n} \frac{x-t_i}{t_i-t_j} (j \neq i)$

La forma en que se define el polinomio de LaGrange hace evidente que el spline que genera es un polinomio de grado k y que es k-veces diferenciable. Esta solución es de gran elegancia matemática aunque tiene algunas desventajas:

El costo computacional es elevado

El tener un polinomio de grado k para colocar k puntos hace que sea muy fácil caer en lo que en Machine Learning se llama sobreestimación, es decir, al tratar de forzar nuestro polinomio a colocar los k puntos se cae en el peligro de dejar de representar la función intrínsica por la cual se generarón estos puntos.

Las variaciones que presenta un polinomio de grado k son exponenciales de grado k al variar un punto en el dominio. Es decir al mover un punto $g(t_{i)}$ (nudo del spline) por un δ_i la variación de la imagen de esta variación varía a su vez en un orden de δ_i elevado a la k

Splines Lineales Por Segmentos

Otra implementación que cumple con nuestra definición de splines llega de manera natural al contruir los splines de manera que S sea una linea recta que una los nudos del spline:

Definimos un spline lineal como la función:

$I_2 g = g(t_i) + (x - t_i)\delta_k, x \in [t_i - ti]$

donde

$\delta_k = g(t_{i+1}) - g(t_i)/(t_{i+1} - t_i)$

definido para cada intervalo $[t_i, t_{i+1}]$

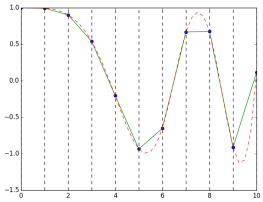


Figure 11: Spline Lineal

Splines Parabólicos

Otra forma de definir splines, ahora con grado 2 es mediante funciones cuadráticas por segmentos. En este caso las parábolas son elegidas para interpolar los puntos entre dos nudos.

Para ello definimos puntos intermedios a cada uno de los intervalos $[t_i, t_{i+1}]$.

Escogemos un punto intermedio a cada intervalo llamado $t_{i+1/2}$ y definimos la funcion de interpolación:

$I_3(x) = g(t_i)\frac{(x-t_{i+1/2})(x-t_{i+1})}{(t_i-t_{i+1/2})(t_i-t_{i+1})} + g(t_{i+1/2})\frac{(x-t_i)(x-t_{i+1})}{(t_{i+1/2}-t_i)(t_{i+1/2}-t_{i+1})} +$
$g(t_{i+1})\frac{(x-t_i)(x-t_{i+1/2})}{(t_{i+1}-t_i)(t_{i+1}-t_{i+1/2})}$
en donde,
$g(t_{i+1/2}) = \frac{g(t_{i+1})-g(t_i)}{2}$

Splines Cúbicos

Los splines cúbicos también llamados splines naturales son la mejora aproximación al problema de la interpolación de valores. Al pertenecer a los polinomios de grado 4 ($\Pi_{\prec 4}$) tienen la bondad de tener 1ª y 2ª derivadas continuas en el intervalo en que se encuentran definidas. De esta forma, intuitivamente, se garantiza la suavidad de la función de interpolación y el mejor ajuste a los datos conocidos.

Hasta el momento la única condición que hemos impuesto, aunque de forma implícita a los splines es que la función:

$S : [a, b] \in C^0$

es decir, que nuestro spline definido parcialmente por polinomios en conjunto sea una función contínua en todo el intervalo. Hay que hacer notar que, como estamos hablando de polinomios, estos son continuos en los intervalos $[t_i, t_{i+1}]$. Por lo anterior es que los únicos puntos en que tenemos que poner condiciones son en los **nudos** o puntos de unión del spline.

Para los splines cúbicos o naturales ponemos algunas condiciones extra:

$S' : [a, b] \in C^1$

$S'' : [a, b] \in C^2$

$S''(t_0) = S''(t_n) = 0$

Posteriormente abundaremos en las propiedades de este tipo de spline por sus características únicas que resultan ser de gran utilidad en la interpolación de valores de una función dada.

Splines de Dimensiones Mayores

Como hemos podido observar en los tipos de splines anteriores, en particular los splines lineales , parabólicos y cúbicos , cada uno de estos tipos está definido dentro de un conjunto o espacio de funciones llamado $\Pi_{\prec n}$. n=2,3,4 respectivamente. A estos espacios de funciones los llamamos **espacio de polinomios de orden n**. Es decir, el conjunto de polinomos de grado menor a n. Es fácil de demostrar que este espacio forma un un espacio vectorial de forma que sus conbinaciones lineales también pertenecen al mismo espacio. Ésta es una notación favorable para la definición de splines.

Definición.

$\Pi_{\prec n,\xi}$ Denotará al espacio de polinomios de grado n definido en el conjunto de intervalos

$\xi = \{[t_i, t_{i+1}]|ti, ti + 1] \subset [a, b], a = t0 < t1 < ... < tn = b\}]$

Es posible demostrar, aunque no se hará en este documento, que este espacio es un espacio vectorial en el cual todo spline de orden n con nodos definidos por ξ se puede definir como una combinación lineal de splines

$S = \alpha_i \varphi_i$

B-Splines

Basados en la idea de un espacio vectorial que pueda generar splines mediante combinaciones lineales surge la idea de los B-Splines o Basis Splines siguiendo la idea de crear una base para el espacio vectorial $\Pi_{\prec n,\xi}$. Esta base de funciones se define recursivamente de la siguiente forma:

Dados
$n > 0$
y
$a = t_0 < t_1 < ... < t_n = b$
definimos

$N_{1,i}\{^{1, t_i < x < t_{i+1}}_{0, d.o.f.}$
$N_{n,i} = N_{n-1,i}\frac{x-t_i}{t_{i+n-1}-t_i} + N_{n-1,i+1}\frac{t_{i+n}-x}{t_{i+n}-t_{i+1}}$

como el conjunto base del espacio vectorial $\Pi_{\prec n,\xi}$

Una de las ventajas de los B-Splines es que generan una base relativamente pequeña de splines. De esta forma es posible realizar regresión por Splines, la cual consiste en métodos de regresión lineal sobre una base de splines (no lineales) aumentando así las aproximaciones de este tipo.

La desventaja a su vez, de este método, es que no asegura la colocación de los puntos deseados (puntos conocidos de la función g) sino una aproximación de estos.

7 Características

Los splines se pueden caracterizar, como hemos visto con ciertas propiedades que los definen:

Orden del spline. Nos referimos al orden de un spline como el mayor grado de los polinomios que lo forman mas 1. Es decir, decimos que un spline es de orden n si todos sus polinomios tienen grado ¡ n.

Número de nudos. El número de nudos o de intervalos que particiónan el dominio del spline nos permite en cierta forma definir la precisión o definición del spline que se generará. Independientemente del orden del spline, mientras más nudos tenga el spline mayor será la precisión con que interpolará puntos.

Distribución de los nudos. Adicionalmente al número de los nudos, también la forma en que éstos están distribuidos en el dominio del spline

8 Splines Cúbicos

Introducción

Como ya se mencionó previamente, cuando tenemos un conjunto de $n + 1$ puntos en el intervalo $[a, b]$ y queremos colocar una curva que cumpla con ciertas características, una forma de hacerlo es mediante el uso de "splines", los cuales

son un tipo de combinación de polinomios que se van uniendo en los $n+1$ puntos que tenemos $((x_0, y_0), (x_1, y_1), ..., (x_n, y_n))$ lo que tendremos n polinomios a lo largo de nuestro intevalo.

En este apartado veremos un tipo muy particular de splines que son los naturales o cúbicos. Como es de intuirse, los polinomios que los formarán son de grado 3 y por lo tanto cada uno de los n polinomios que se formen tendrá la siguiente forma:

$$a + bx + cx^2 + dx^3 \tag{1}$$

Así entonces necesitamos hallar cada uno de los cuatro coeficientes de cada uno de los n polinomios. Para esto definamos:

$$S_{i-1,i}(x) : \text{El polinomio del intervalo } [i-1, i]$$

$$a_{i-1,i}; b_{i-1,i}; c_{i-1,i}; d_{i-1,i} : \text{Los coeficientes del polinomio } S_{i-1,i}(x)$$

Por lo tanto, el i-ésimo polinomio estará dado por:

$$S_{i-1,i}(x) = a_{i-1,i} + b_{i-1,i}x + c_{i-1,i}x^2 + d_{i-1,i}x^3 \tag{2}$$

La tarea de hallar los coeficientes la haremos por medio de sistemas de ecuaciones que se van generando con las condiciones que se necesitan cumplir. De entrada para cada uno de lo polinomios tenemos dos ecuaciones que se tienen que cumplir para que se mantenga el sentido:

$$a_{i-1,i} + b_{i-1,i}x_{i-1} + c_{i-1,i}x_{i-1}^2 + d_{i-1,i}x_{i-1}^3 = y_{i-1} \tag{3}$$

$$a_{i-1,i} + b_{i-1,i}x_i + c_{i-1,i}x_i^2 + d_{i-1,i}x_i^3 = y_i \tag{4}$$

Continuidad en la primera derivada

Ahora bien, como ya vimos anteriormente, una de las propiedades de los "splines"en general es que el conjunto de polinomios mantenga la suavidad entre vecinos y esto se traduce a:

$$S'_{i-1,i}(x_i) = S'_{i,i+1}(x_i) \tag{5}$$

Por lo que ya tenemos otra ecuación para agregar.

Continuidad en la segunda derivada

Podemos generar una más cuando recordamos que los splines también requieren que la concavidad se mantenga entre polinomios vecinos:

$$S''_{i-1,i}(x_i) = S''_{i,i+1}(x_i) \tag{6}$$

Curvatura Mínima

Cuando generemos todas las ecuaciones que se han comentado para cada polinomio, lo que vamos a tener en total serán $4n$ incógnitas pero $4n-2$ condiciones. Por lo tanto se tiene que generar un par extra para poder resolver el

sistema. Por lo general estas dos condiciones adicionales está basadas en dar valores a la segunda derivada del spline evaluada en los puntos extremos (x_0 y x_n). Lo que hace muy particular a los splines naturales es que el valor que se le da a ambos casos es igual a cero:

$$S''(x_0) = S''(x_n) = 0 \tag{7}$$

Esta condición parece arbitraria. Sin embargo ésta es la que garantiza que la interpolación es una función doblemente diferenciable y continua que además cumple con que aproxima a los puntos con la menos "curvatura" que es posible.

Demostración:

Antes que nada sería bueno recordar que una buena aproximación para medir la *curvatura* de una función $f(x)$ en el intervalo $[a, b]$ es mediante la integral:

$$\int_a^b [f''(x)]^2 \tag{8}$$

Ahora bien, sea $S(x)$ el spline de colocación con las propiedades señaladas; Sea $g(x)$ cualquier otra función de colocación. También hagamos a $f(x)$ la función que se quiere aproximar.

Tenemos:

$$\int_a^b [g''(t) - S''(t)]^2 dt = \int_a^b [g''(t)]^2 dt - 2\int_a^b [g''(t) - S''(t)]S''(t)dt - \int_a^b [S''(t)]^2 dt \tag{9}$$

Dado que $a = x_0 < x_1 < ... < x_n = b$,podemos expresar la integral del segundo término a la derecha del símbolo "=" (al cual denotaremos como D) como:

$$D = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} S''(t)[g''(t) - S''(t)]dt \tag{10}$$

Aplicando integración por partes:

$$D = \sum_{i=0}^{n-1} \{[g'(t) - S'(t)]S''(t)|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} [g'(t) - S'(t)]S'''(t)dt\} \tag{11}$$

Ya que $S'''(t)$ sobre el intervalo $[x_i, x_{i+1}]$ es constante (llamémosle α_i) podemos escribir lo anterior (11) como:

$$\int_{x_i}^{x_{i+1}} [g'(t) - S'(t)]S''(t)dt = \alpha_i [g(t) - S(t)]|_{x_i}^{x_{i+1}} \tag{12}$$

Sin embargo, recordemos que:

$$g(x_i) = f(x_i) = S(x_i)\ \forall_{i=1,2,\ldots,n}$$

Por lo que:

$$\alpha_i[g(t) - S(t)]|_{x_i}^{x_{i+1}} = \alpha_i[g(x_{i+1}) - S(x_{i+1}) - (g(x_i) - S(x_i)]$$

$$= \alpha_i[(f(x_{i+1}) - f(x_{i+1})) - (f(x_i) - f(x_i)]$$

$$= \alpha_i[(0) - (0)]$$

$$= 0$$

De ahí que la expresión se reduce a,

$$D = \sum_{i=0}^{n-1}\{S''(x_{i+1})[g'(x_{i+1}) - S'(x_{i+1})] - S''(x_i)[g'(x_i) - S'(x_i)]\}$$

$$= S''(x_n)[g'(x_n) - S'(x_n)] - S''(x_{n-1})[g'(x_{n-1}) - S'(x_{n-1})] +$$

$$S''(x_{n-1})[g'(x_{n-1}) - S'(x_{n-1})] - S''(x_{n-2})[g'(x_{n-2}) - S'(x_{n-2})] +$$

$$\ldots$$

$$S''(x_2)[g'(x_2) - S'(x_2)] - S''(x_1)[g'(x_1) - S'(x_1)] +$$

$$S''(x_1)[g'(x_1) - S'(x_1)] - S''(x_0)[g'(x_0) - S'(x_0)]$$

De lo cual, al cancelar términos, sólo quedan los de los valores extremos:

$D = S''(x_n)[g'(x_n) - S'(x_n)] - S''(x_0)[g'(x_0) - S'(x_0)]$

Pero como ya habíamos comentado sabemos que se cumple que

$S''(x_n) = S''(x_0) = 0,$

por lo que resolvemos que:

$$D = 0 \tag{13}$$

De lo anterior que nuestra primera igualdad (9) queda como:

$$\int_a^b [g''(t) - S''(t)]^2 dt = \int_a^b [g''(t)]^2 dt + 0 - \int_a^b [S''(t)]^2 dt \tag{14}$$

Despejando $\int_a^b [g''(t)]^2 dt$,

$$\int_a^b [g''(t)]^2 dt = \int_a^b [g''(t) - S''(t)]^2 dt + \int_a^b [S''(t)]^2 dt \tag{15}$$

Pero sabemos que

$$\int_a^b [g''(t) - S''(t)]^2 dt \geq 0 \tag{16}$$

$$\implies \int_a^b [g''(t) - S''(t)]^2 dt + \int_a^b [S''(t)]^2 dt \geq \int_a^b [S''(t)]^2 dt \tag{17}$$

$$\implies \int_a^b [g''(t)]^2 dt \geq \int_a^b [S''(t)]^2 dt \tag{18}$$

Y como vimos al principio de la demostración, lo que tenemos en la última desigualdad es cualquier $g(x)$ tendrá una mayor curvatura que la $S(x)$ que se construya bajo las propiedades de los splines cúbicos, es decir, agregando aquellas dos últimas condiciones al sistema de ecuaciones:

$$S''(x_0) = 0 \wedge S''(x_n) = 0 \tag{19}$$

Y con esto podemos concluir que cualquier $S(x)$ garantiza la **curvatura mínima**.

Comparativa con otros modelos

Cuando se contrasta a los splines cúbicos con cualquier otro vemos que los naturales llegan a ser los mejores pues satisfacen todas las condiciones de los splines con el mínimo grado polinomial.

No bastando con ello, los splines naturales consiguen generar una función que garantiza ser la que contiene menor curvatura al colocarse sobre los puntos que tenemos en el intervalo $[a, b]$, por lo que en sí, la complejidad de la función es la menor entre todas las que podríamos generar.

9 Demostración Uso de Splines Naturales y Ventanas Móviles

Para ejecutar la aplicación, se requiere tener instalada la consola R o RStudio en el equipo, si no se tiene instalado, se puede descargar de los sitios

`http://cran.itam.mx/` o `https://www.rstudio.com/products/RStudio/#Desktop`

Dentro se encuentra la distribución acorde a su sistema operativo.

Una vez dentro de la consola de R se deben ejecutar la siguientes instrucciones.

library(shiny) runGitHub("CompuStat", username = "eduardomtz", subdir = "SplinesNaturales")

La primera instrucción es debido a

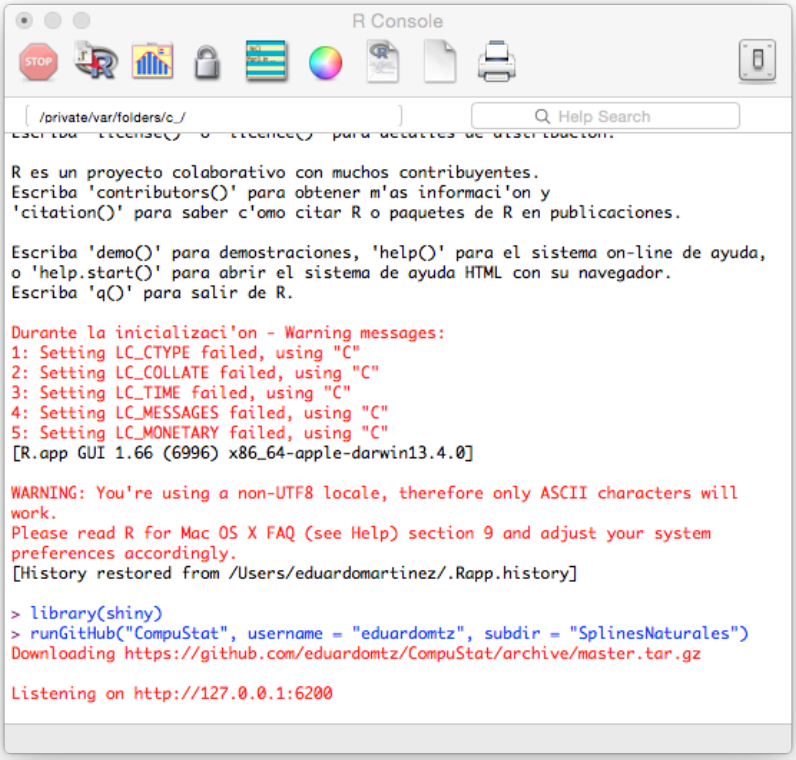
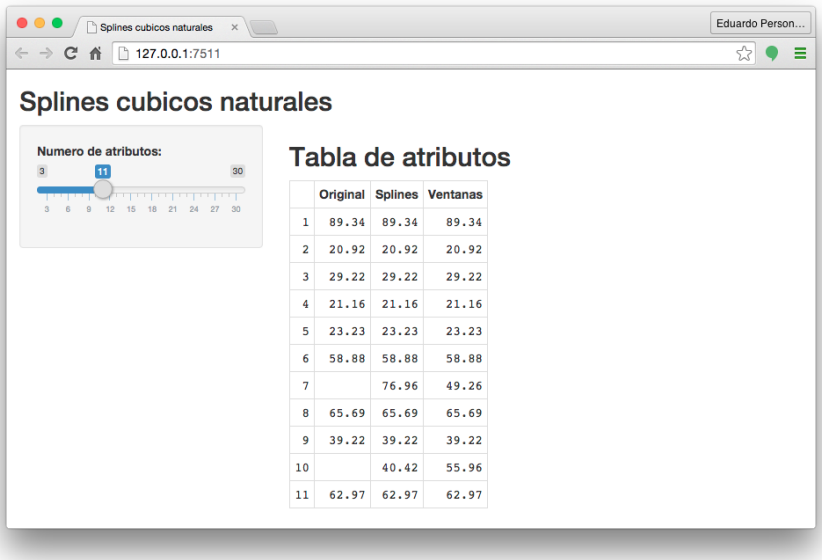


Figure 12: Consola R

que la aplicación utiliza un paquete especifico de R llamado Shiny, y la segunda instrucción descarga la aplicación de GitHub y la ejecuta. Una vez realizados estos pasos, se podrá observar la aplicación en el navegador predeterminado.



	Original	Splines	Ventanas
1	89.34	89.34	89.34
2	20.92	20.92	20.92
3	29.22	29.22	29.22
4	21.16	21.16	21.16
5	23.23	23.23	23.23
6	58.88	58.88	58.88
7		76.96	49.26
8	65.69	65.69	65.69
9	39.22	39.22	39.22
10		40.42	55.96
11	62.97	62.97	62.97

Figure 13: Visualizando aplicacion

La aplicación genera números aleatorios de acuerdo al número de atributos que se elijan en el slide, en un intervalo de 1 a 100. Una vez generados los datos, agrega huecos en los datos aleatoriamente, para aproximadamente un 20% de los datos generados.

El primero y el último dato no pueden ser huecos, debido a que estamos realizando un proceso de interpolación, mediante splines y ventanas móviles.

Una vez generados los datos y los huecos se ejecutan los algoritmos para generar un spline cubico natural, y se genera también una ventana móvil con promedios en los 2 datos anteriores y 1 dato posterior, para calcular el dato faltante.

En la gráfica se muestran ambos métodos, splines con cruces azules y una linea en color verde que representa el modelo completo del spline y para ventanas móviles se representa el dato calculado mediante una X roja.

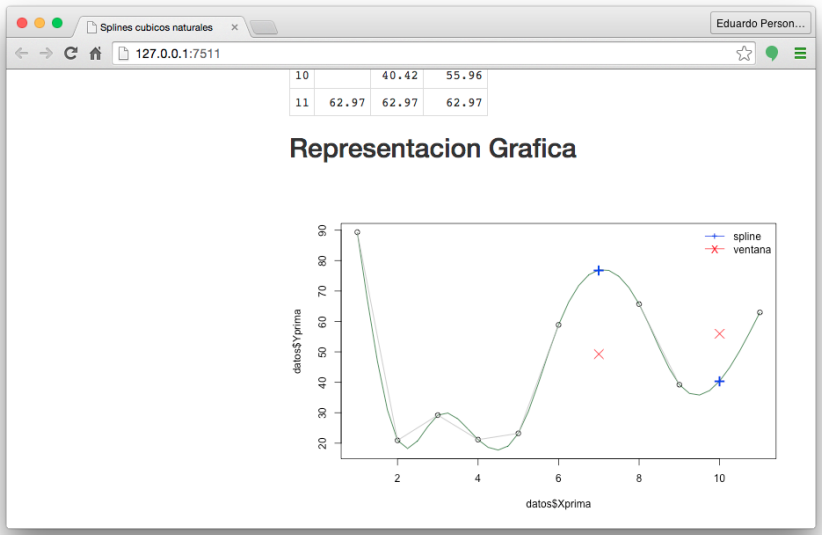
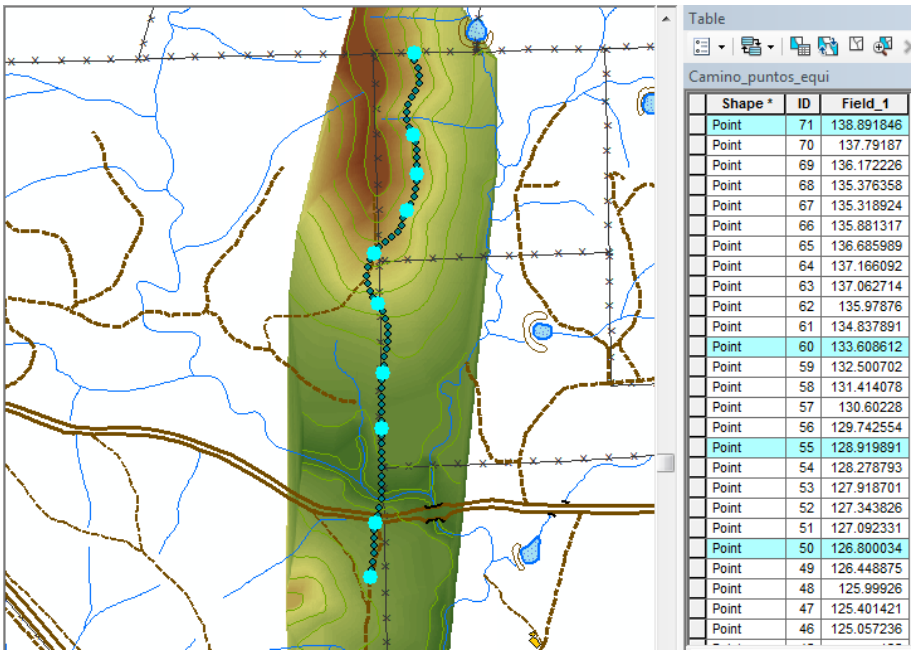


Figure 14: Visualizando aplicacion

10 Uso de splines naturales para cálculo de perfil en caminos

Se realiza también un caso de uso de splines, utilizando información geográfica, se consideran ciertos puntos de un camino, los cuales tienen altura definida por el relieve del terreno.



Shape *	ID	Field_1
Point	71	138.891846
Point	70	137.79187
Point	69	136.172226
Point	68	135.376358
Point	67	135.318924
Point	66	135.881317
Point	65	136.685989
Point	64	137.166092
Point	63	137.062714
Point	62	135.97876
Point	61	134.837891
Point	60	133.608612
Point	59	132.500702
Point	58	131.414078
Point	57	130.60228
Point	56	129.742554
Point	55	128.919891
Point	54	128.278793
Point	53	127.918701
Point	52	127.343826
Point	51	127.092331
Point	50	126.800034
Point	49	126.448875
Point	48	125.99926
Point	47	125.401421
Point	46	125.057236

Figure 15: Puntos considerados en camino

Se generó un spline natural a partir de puntos conocidos en el camino, y a partir de este modelo, se interpolaron las alturas restantes.

Con el objetivo de lograr un camino suavizado, debido a la propiedad de los splines naturales de curvatura mínima.

En la gráfica se puede observar en gris la forma del relieve o terreno natural y en verde se muestra un camino calculado a partir de splines naturales tomando como base los puntos a las distancias 10, 80, 200, 270, 360, 430, 500, 550, 600 y 710 metros.

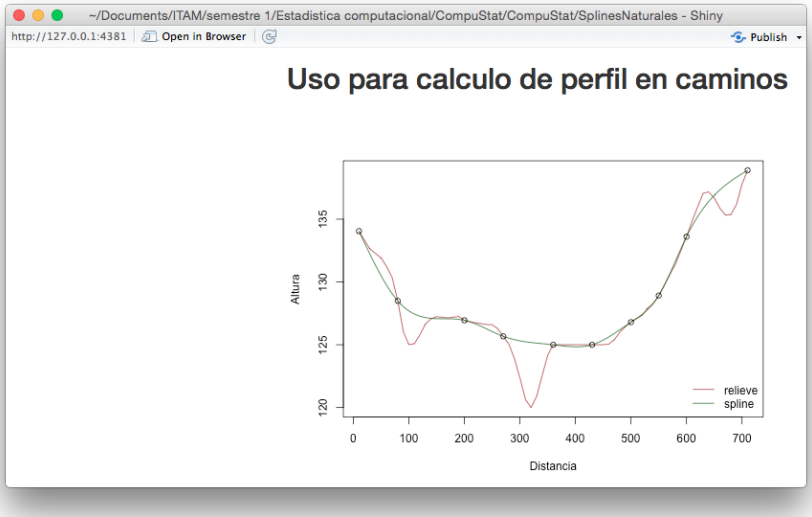


Figure 16: Perfil de camino con splines naturales

11 Referencias

Kuri-Morales, A. "Natural Splines." Instituto Tecnológico Autónomo de México.

Parnell, C. An Introduction to the Approximation Functions. Apuntes de Anlisis Numrico. Captulo 3. Recuperado de:
$http://www-solar.mcs.st-andrews.ac.uk/\ clare/Lectures/num-analysis/Numan_chap3.pdf$

Calvo, N. 2011. Computacin grfica. Universidad Nacional del Litoral, Facultad de Ingenieras y Ciencias Hdricas. Recuperado de:
$http://www.cimec.org.ar/\ ncalvo/interpolacion_doc.pdf$

Wayman, J. C. 2003. Multiple Imputation for Missing Data: What is it and how can I use it? Annual Meeting of the American Educational Research Association. Recuperado de:
$http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf$

Polynomial interpolation. (2015, September 5). In Wikipedia, The Free Encyclopedia. Retrieved 01:36, September 23, 2015, from
$https://en.wikipedia.org/w/index.php?title=Polynomial\_interpolation\&oldid=$

679644573

Perfil topográfico. (2015, 1 de junio). Wikipedia, La enciclopedia libre. Fecha de consulta: 03:07, septiembre 23, 2015 desde $https://es.wikipedia.org/w/index.php?title=Perfil\_topogr\%C3\%A1fico\&oldid=$ 82878037

$http://www.cs.mtu.edu/\ shene/COURSES/cs3621/NOTES/spline/B-spline/bspline-basis.html$

Ruedin, Didier. Moving average to fill holes (interpolation) publicado en sitio $http://druedin.com/2013/06/15/moving-average-to-fill-holes-interpolation/$

de Boor, A Practical Guide to Splines (revised edition) (2001) Springer Applied Mathematical Sciences Vol. 27, 2001