Servy, Elsa
Cuesta, Cristina
Marí, Gonzalo
Armida, María Luz
Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística

UTILIZACION DEL PAQUETE "KERNSMOOTH" DE R PARA CONSTRUIR SUAVIZADOS LOESS Y BANDAS DE VARIABILIDAD A DATOS DE LA ENCUESTA DE OCUPACION HOTELERA

1. INTRODUCCIÓN

Los modelos de regresión constituyen una herramienta muy útil y potente para los analistas estadísticos cuando su objetivo es relacionar una variable respuesta con una o más variables explicativas. A partir de ellos se pueden estimar parámetros (que a menudo gozan de una explicación plausible en términos del problema), realizar inferencias y hacer predicciones. Sin embargo estos modelos tienen una estructura rígida (lineal, cuadrática, etc.) que a menudo no logra "captar" el comportamiento de los datos en todo el campo de variación de las variables explicativas. Por otro lado las inferencias realizadas sobre dichos modelos sólo son válidas bajo rigurosos supuestos distribucionales. Si la realidad no se ajusta a esta situación debe recurrirse a otras herramientas que permitan tanto representar adecuadamente los datos como flexibilizar los supuestos. Entre estas herramientas se encuentran los denominados "suavizados" que no suponen de antemano que las variables (explicativas y respuesta) estén relacionadas en una forma particular y cuyo resultado son curvas (en el caso univariado) "suaves" que representan adecuadamente los datos. Un tipo particular de suavizado es el constituido por las curvas "lowess" (Locally Weighted Scatterplot Smoothing) o más genéricamente "loess" (local regression) desarrolladas por Cleveland (1979) y que se basan en ajustar modelos de regresión polinómicos locales para estimar cada punto y luego unir las estimaciones. En un primer paso se define el ancho de la ventana que encierra los puntos cercanos (o vecinos) al punto que se quiere estimar. Luego se elige una función de ponderación que le dé mayor peso a las observaciones más cercanas y menor peso a las observaciones más alejadas (dentro de la ventana). Usando mínimos cuadrados ponderados se ajusta una regresión polinómica dentro de la ventana y se estima el punto en cuestión. Estos pasos se repiten para cada observación en el conjunto de los datos y/o para otros puntos dentro del campo de variación de la variable explicativa. Los valores estimados por estas regresiones se grafican en el diagrama de dispersión y se unen produciendo una curva de regresión no paramétrica.

Como la búsqueda de la curva es inductiva, hay muchas curvas posibles. La elección debe proveer una curva que no sea demasiado "suave" ni demasiado "rugosa". El suavizado de la curva tiene una relación directa con el grado del polinomio, la función de ponderación y con el ancho de la ventana. La elección de la ventana puede realizarse de dos maneras. La primera es trabajar con ventanas de un ancho cuya amplitud está prefijada por el analista ("bandwidth"), otra manera de seleccionar las ventanas es por el método del vecino más cercano eligiendo la ventana de modo que cada punto a estimar contenga una proporción especificada de las observaciones ("span"). El suavizado resultante de ajustar una curva de regresión polinómica local a los datos muestrales puede ser acompañado, en el gráfico, por

unas bandas de variabilidad construidas por el método de bootstrap para dar una idea de la precisión de la estimación.

El ajuste de las curvas, a partir de regresión polinómicas locales, ha tenido gran desarrollo en los programas de computación durante los últimos años. La flexibilidad y aplicabilidad de estos programas son las características que determinan la preferencia de los usuarios. Algunos programas ofrecen mayor versatilidad que otros en cuanto a la posibilidad de escoger un determinado grado del polinomio, función de ponderación y tipo de ventana seleccionada.

Este trabajo tiene por objetivo presentar el ajuste de curvas loess utilizando el paquete Kernsmooth de R (software de distribución libre) describiendo sus principales características.

Se analizarán datos provenientes de la Encuesta de Ocupación Hotelera (EOH) que el Instituto Nacional de Estadística y Censos realiza mensualmente en distintas localidades de nuestro país. En particular se mostrarán ajustes para relacionar las tasas de ocupación de habitaciones, tasas de ocupación de plazas y plazas por personal ocupado para la ciudad de Rosario y Mar del Plata para el año 2005.

2. METODOLOGÍA

La relación entre una variable explicativa y una respuesta puede expresarse a través del modelo: $y_i = m(x_i) + \varepsilon_i$, donde la curva de regresión $m(x)$ es la esperanza condicional $m(x) = E(Y / X = x)$. Cuando un modelo de regresión paramétrico no es apropiado (porque no llega a "captar" adecuadamente la estructura que relaciona ambas variables), una alternativa es utilizar un modelo de regresión no paramétrica (que elimina la restricción paramétrica sobre $m(x)$)

Por definición,

$$m(x) = E(Y / X = x) = \int y\, f(y/x)\, dy = \int y \frac{f(x,y)}{f_X(x)} dy \qquad (2.1)$$

donde $f_X(x)$ es la función de densidad marginal de X, $f(x,y)$ es la función de densidad conjunta de X e Y. Si cada una de estas funciones de densidad es estimada en forma no paramétrica a través del método de estimación mediante nucleos, "Kernel estimation" (Simonoff, 1996), se obtiene el estimador de Nadaraya-Watson:

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} \equiv \sum_{i=1}^{n} w_i y_i \qquad (2.2)$$

donde,

h: es la semiamplitud de la ventana de datos vecinos a x que contribuyen a la estimación, es el parámetro de suavizado que más afecta a la estimación de la curva

K: denominada función de núcleo o ponderación ("Kernel function"), es una función tal que cumple con ciertas propiedades que garantizan que $\hat{f}_X(x)$ y $\hat{f}(x,y)$ sean estimaciones de las funciones de densidad correspondientes. En la Tabla 1 se presentan las funciones de

núcleo más comunmente utilizadas. Se ha mostrado (Simonoff, 1995) que el uso de una u otra función de ponderación no afecta sensiblemente las curvas resultantes y por ello a menudo se prefiere la función gausiana debido a sus propiedades de diferenciabilidad.

Tabla 1. Funciones de núcleo (K) más difundidas

Kernel	Forma
Epanechnikov	$\frac{3}{4}(1-\mathrm{u}^2)$
Biweight	$\frac{15}{16}(1-\mathrm{u}^2)^2$
Triweight	$\frac{35}{32}(1-\mathrm{u}^2)^3$
Normal (Gaussiana)	$(2\pi)^{-1/2}\mathrm{e}^{-\mathrm{u}^2/2}$
Uniforme	$1/2$

El ancho de la ventana (h) puede ser definido de acuerdo a uno de los siguientes conceptos:

- "bandwidth": anchos de ventana pre-fijado por el investigador, no dependen de los valores muestrales, en general son todos de igual amplitud aunque el analista puede definir amplitudes diferentes. Simonoff (1995) presenta varias metodologías para la selección del ancho óptimo: regla gausiana, validación cruzada y el principio "plug-in".
- "span": anchos de ventana construídos a partir de una proporción de los datos muestrales, por ello, los anchos serán variables dependiendo del agrupamiento de los datos.

A partir de (2.2) se observa que $\hat{m}_{NW}(x)$ es una función lineal de y con pesos $w_i = \frac{1}{nh}\frac{K\left(\frac{x-x_i}{h}\right)}{\hat{f}_X(x)}$ y puede deducirse que $\hat{m}_{NW}(x)$ es también la solución por mínimos cuadrados ponderados de β_0 en una regresión donde la variable respuesta es explicada sólo por un intercepto, es decir la expresión de $\hat{\beta}_0$ que minimiza la expresión: $\sum_{i=1}^{n}(y_i-\beta_0)^2 K\left(\frac{x-x_i}{h}\right)$ es $\hat{m}_{NW}(x)$. Por lo tanto, $\hat{m}_{NW}(x)$ corresponde a la aproximación local de $m(x)$ por una constante. Los mayores valores de los pesos de y corresponden a los $x_i s$ cercanos a x.

Este estimador sugiere, entonces, ajustar polinomios locales de mayor grado ya que una constante local sólo tendría sentido sobre un pequeño vecindario (ancho de ventana muy pequeño). Así, el estimador de la regresión polinómica local es el que minimiza:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\ldots-\beta_p(x-x_i)^p\right)^2 K\left(\frac{x-x_i}{h}\right) \qquad (2.3)$$

Como puede advertirse en la expresión (2.3), la estimación dependerá del grado del polino-

mio, de la función K de ponderación y del ancho de ventana h.

La matriz de diseño para estimar y en el punto x es $X_x = \begin{pmatrix} 1 & x-x_1 & \dots & (x-x_1)^p \\ 1 & x-x_2 & \dots & (x-x_2)^p \\ \vdots & \vdots & & \vdots \\ 1 & x-x_n & \dots & (x-x_n)^p \end{pmatrix}$ y la matriz de los pesos es $W_x = \frac{1}{h} diag\left[K\left(\frac{x-x_1}{h}\right),\dots,K\left(\frac{x-x_2}{h}\right)\right]$. Luego si $X_x' W_x X_x$ es invertible, el estimador del vector β resultante es:

$$\hat{\beta} = \left(X_x' W_x X_x\right)^{-1} X_x' W_x y \qquad (2.4)$$

A partir de (2.4) se obtiene $\hat{m}(x) = e_1' \left(X_x' W_x X_x\right)^{-1} X_x' W_x y$ donde e_1' es un vector fila con (p+1) columnas que contiene el valor 1 en la primera y 0 en las restantes.

Una vez realizada la estimación para todos los valores de x (e incluso para otros valores que pertenezcan al campo de variación de la variable explicativa, para lograr una curva más suave), estas se unen formando la curva suavizada.

A continuación se puede construir un "Gráfico de Variabilidad" que consiste en acompañar la curva estimada con otras dos curvas que la "envuelven" y muestran un grado de certeza de $\hat{m}(x)$. Una forma de construir estas curvas (o bandas) del gráfico de variabilidad, es a través de una metodología de remuetreo (bootstrap). El Bootstrap consiste fundamentalmente en tratar la muestra como si fuese la población y aplicar un muestreo con reposición para generar una estimación empírica de la distribución muestral del estadístico (en este caso $\hat{m}(x)$) . Al ser una técnica no paramétrica, el Bootstrap tiene la ventaja de que no precisa conocer la función de distribución teórica de los datos. Las bandas de variabilidad se construyen seleccionando (repetidamente) muestras con reemplazo de los datos originales y reajustando las curvas loess para cada nueva muestra. Luego se calcula el percentil del 2.5% y del 97.5% a través de todas las muestras y se unen para todos los valores de x (Härdle (1990), Fox(2000)) . Las curvas resultantes no corresponden a un intervalo de confianza del 95% ya que en su construcción se está ignorando el sesgo en la estimación de la curva. Unas "bandas" angostas (o cercanas a $\hat{m}(x)$) sugieren una mayor estabilidad en las estimaciones. Si se utiliza una función de núcleo Gausiana, estas bandas se ensanchan en los extremos del campo de variación de x .

3. EL PAQUETE KERNSMOOTH DE R

La plataforma R fue desarrollada en 1992 por Ross Ihaka y Robert Gentleman, de la Universidad de Aukland, como una implementación "abierta" del programa S-PLUS. Está constituida por un lenguaje y ambiente computacional de fácil acceso. Lo conforman diferentes "paquetes" que proveen las herramientas para realizar los distintos análisis y gráficos estadísticos. Si bien hay varios paquetes que se utilizan para el cálculo de las estimaciones de regresiones polinómicas locales, el utilizado para analizar los datos en este trabajo fue el denominado "Kernsmooth" desarrollado por Wand (1995) y utilizado por Simonoff (1995) para aplicar la metodología presentada en la sección 2. En el paquete Kernsmooth, a través de la función *locpoly*, se estima una curva de regresión usando polinomios locales (cuyo

grado está definido por el usuario y puede ser 0,1,2…), con anchos de ventana prefijados por el analista (o determinado por un método plug-in a través de la función *dpill*) pero no permite definir el ancho de ventana como "span" y con una función de ponderación de gausiana. En la Sección 4 se presentan algunos de estos programas.

4. APLICACIÓN

Hacia fines del año 2003 la Secretaría de Turismo de la Nación (SECTUR) y el Instituto Nacional de Estadística y Censos (INDEC) firmaron un convenio para medir el impacto y la participación del turismo en el conjunto de la economía de la Argentina. Entre los operativos diseñados para tal fin se encuentra la Encuesta de Ocupación Hotelera (EOH) cuyo objetivo es medir el impacto del turismo internacional e interno sobre la actividad de los establecimientos hoteleros y para-hoteleros, la oferta y utilización de infraestructura, la evolución de tarifas, etc.. En el año 2004 la EOH se desarrolló con periodicidad mensual en 17 localidades, mientras que en el año 2005 se redefinió en 39 localidades de 6 regiones turísticas del país (determinadas por SECTUR).

Entre las principales variables investigadas se encuentra la categoría del establecimiento, cantidad de personal ocupado, habitaciones o unidades y plazas disponibles, entrada de viajeros según lugar de residencia, tarifa promedio, etc. Se define como establecimientos "hoteleros" a aquellos categorizados como 1,2,3,4 y 5 estrellas y apart-hoteles, mientras que el grupo de establecimientos "para-hoteleros" lo conforman los hoteles sindicales, albergues, cabañas, bungalows, hospedajes, bed & breakfast, hosterías, residenciales, etc.

Los principales índices estudiados son: la tasa de ocupación de plazas (TOP), la tasa de ocupación de habitaciones (TOH) y las plazas por personal ocupado (PPO) que se definen como:

$$TOP = \frac{plazas\ ocupadas}{plazas\ disponibles \times días\ abiertos} \times 100$$

$$TOH = \frac{(unidades + habitaciones\ ocupadas)}{(unidades + habitaciones\ disponibles) \times días\ abiertos} \times 100$$

$$PPO = \frac{plazas\ disponibles}{personal\ ocupado}$$

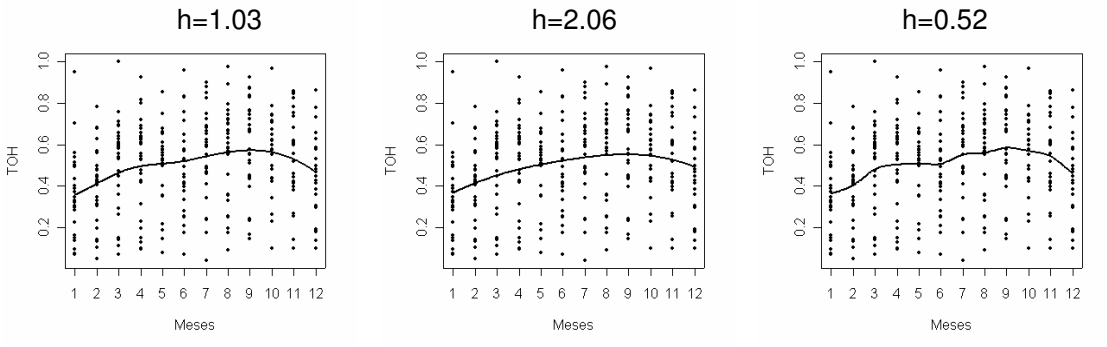
En este trabajo se analizan, con la metodología propuesta, datos correspondientes a las ciudades de Rosario y Mar del Plata, del año 2005.

Rosario

La Ciudad de Rosario fue incluída en la muestra como dominio de estudio a partir del año 2005 y debido a la escasa cantidad de establecimientos para-hoteleros, estos no fueron considerados en el análisis. En la Figura 1 se muestra la Tasa de Ocupación de Habitaciones por mes ajustada por regresiones polinómicas locales con diferentes anchos de ventana (h=1.03 h=2.06 y h=0.52). El primero de los anchos de ventana presentados corresponde al óptimo obtenido por un procedimiento plug-in. Los otros dos anchos de ventana corresponden al doble y a la mitad del óptimo mostrando la influencia de h sobre la estimación de la curva. En todos los casos el grado del polinomio considerado fue 1 y se utilizó una función de núcleo gausiana (que es la disponible en el paquete). Se observa claramente una curva creciente de enero a diciembre (con una leve depresión en junio) y que decae a partir de octubre. Cuando la amplitud de la ventana es mayor a la óptima, se observa una curva so-

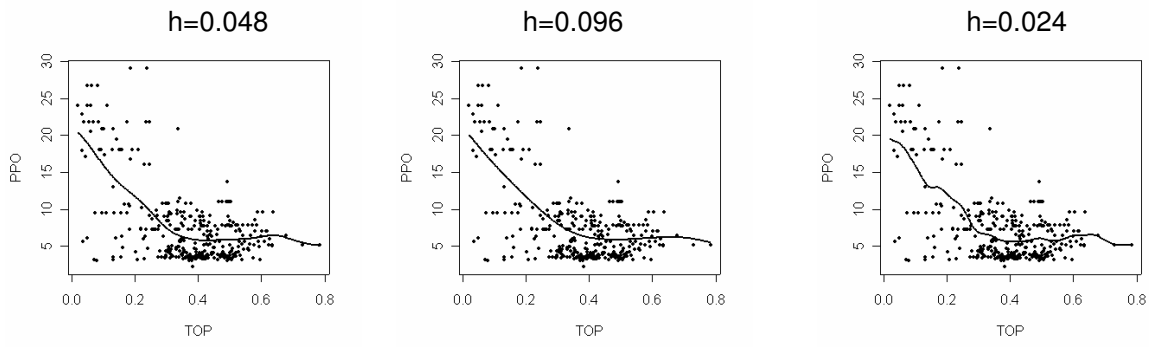
bre-suavizada mientras que cuando el ancho de ventana es menor al óptimo, la curva está sub-suavizada.

Figura 1. Tasa de Ocupación de Habitaciones por mes para la ciudad de Rosario, 2005.



La Figura 2 muestra la comparación del ajuste de Plazas por Personal Ocupado versus la Tasa de Ocupación de Plazas ajustada con diferentes anchos de ventana, el óptimo, el doble y la mitad del óptimo, y con grado del polinomio igual a 1. Nuevamente se observa una sobre-suavización a medida que el ancho de ventana aumenta. Si h aumentara hasta ser igual al rango de variación de TOP, la curva resultante sería la recta obtenida por mínimos cuadrados. Estos gráficos sugieren que cuando los establecimientos están ocupados en el promedio de su capacidad, la cantidad de personal disponible por pasajero está entre 5 y 7. Cuanto más baja es la ocupación hotelera, mayor es la disponibilidad de personal por pasajero (probablemente se trate del personal estable del establecimiento). Por otro lado, cuando la tasa de ocupación de plazas aumenta al punto máximo, la cantidad de personal se estabiliza alrededor de 6 (lo que indicaría que se contrata personal temporario para mejor atención de los pasajeros)

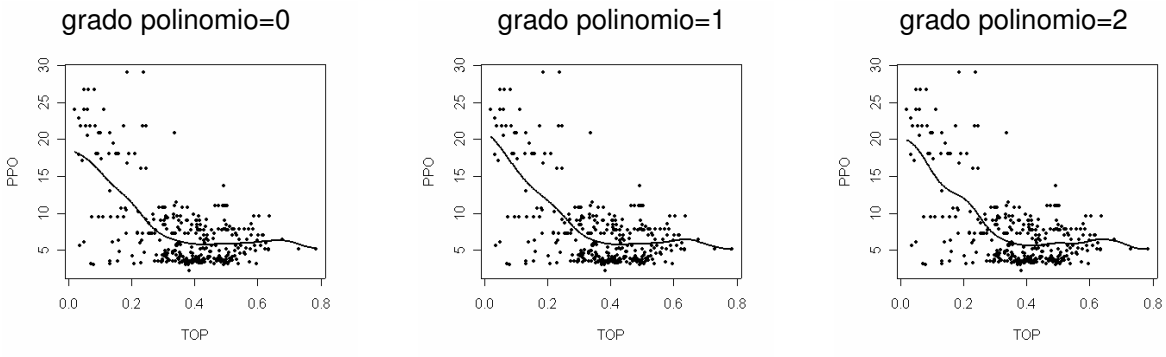
Figura 2. Plazas por Personal Ocupado vs. Tasa de Ocupación de Plazas para la ciudad de Rosario, 2005.



En la Figura 3. Se muestran las curvas de regresión polinómicas locales que se obtienen al relacionar las variables PPO versus TOP con el ancho de ventana óptimo (h=0.048). En este caso se muestra la influencia del grado del polinomio sobre el ajuste, un polinomio de grado 0 corresponde a la estimación de Nadaraya-Watson, el polinomio de grado 1 es el que presenta por defecto el paquete Kernsmooth. Puede observarse que a medida que el grado del polinomio aumenta, la curva se adapta mas a los cambios locales (es más "rugosa"). Varios autores han sugerido la utilización de polinomios de grado impar debido a que

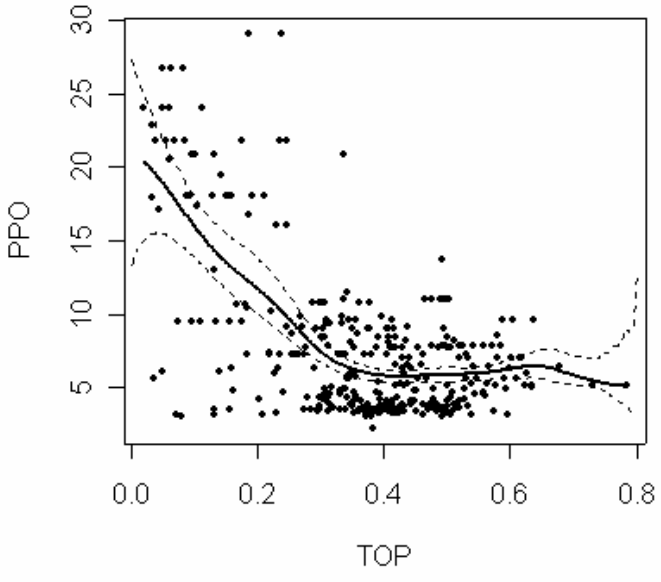
estos reducen el sesgo en la estimación.

Figura 3. Plazas por Personal Ocupado vs. Tasa de Ocupación de Plazas para la ciudad de Rosario, 2005.



En la Figura 4 se estudia la relación entre la Tasa de Ocupación de Plazas versus las Plazas por Personal ocupado acompañadas de las bandas de variabilidad considerando un ajuste por regresión polinómica local con ancho de ventana h=0.048, grado de polinomio 1 y función de núcleo Gausiana. Para la construcción de las bandas se tomaron 2000 muestras con reemplazo de tamaño 330 (cantidad de datos de la muestra original). Las bandas son más angostas en los sectores donde hay más datos en la nube de puntos, garantizando en esa zona un mejor ajuste, y más ancha en los sectores donde hay menor concentración de datos. Hacia los extremos del gráfico las bandas aumentan mucho en su ancho, lo cual es una característica de estos gráficos especialmente cuando se utiliza una función Kernel gausiana. En términos generales, este gráfico remarca la bondad de la curva ajustada.

Figura 4. Gráfico de Variabilidad para el ajuste de Plazas por Personal Ocupado vs. Tasa de Ocupación de Plazas para la ciudad de Rosario, 2005.



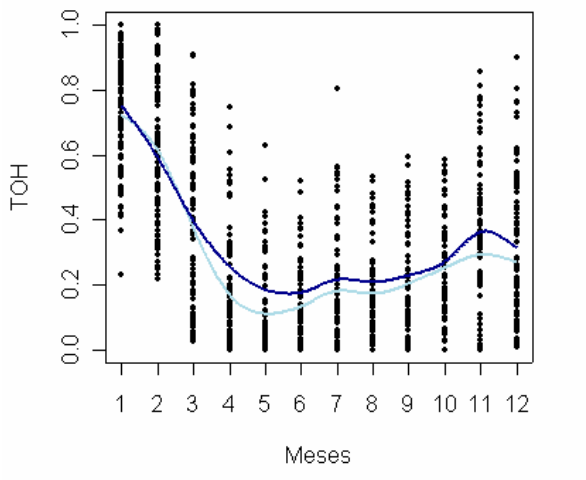
La construcción de las bandas de variabilidad (Figura4) en R se realizó con el siguiente programa:

```
library(KernSmooth)
varplot<-matrix(NA,401,2000)
for (i in 1:2000){
vardata<-cbind(top,ppo)[sample(330,replace=T),]
varplot[,i]<-locpoly(vardata[,1],vardata[,2],bandwidth=0.04803643,range.x=c(0,0.8))$y}
for (i in 1:401) varplot[i,]<-sort(varplot[i,])
plot(top,ppo,cex=0.7,xlab="",ylab="",pch=20,ylim=(2.30,30))
lines(locpoly(top,ppo,bandwidth=0.04803643))
lines(c(0:400)*0.002,varplot[,1950],lty=2)
lines(c(0:400)*0.002,varplot[,50],lty=2)
```

Mar Del Plata

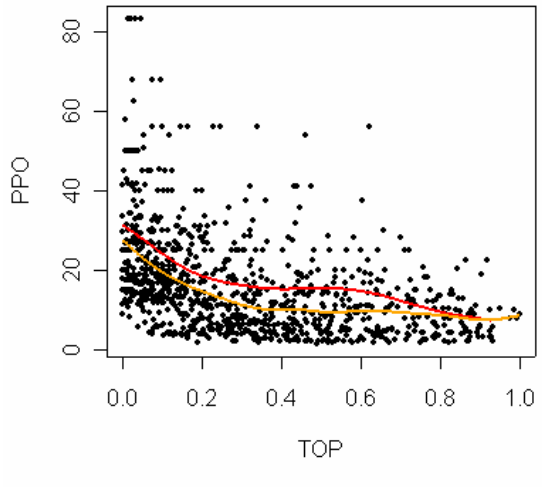
Para la ciudad de Mar del Plata, es de gran importancia la comparación entre establecimientos hoteleros y parahoteleros. Por ello, en la Figura 5 se muestra la Tasa de ocupación de Habitaciones por mes tanto para establecimientos Hoteleros como para Establecimientos Para-hoteleros. El ajuste se realizó a través de regresiones polinómicas locales, con polinomios de grado 1, ancho de ventana óptimo para cada grupo y función de núcleo gausiana. Ambas líneas muestran una gran ocupación para los meses de verano, descendiendo hacia el invierno, donde hay una leve suba en julio y luego un nuevo aumento en noviembre y diciembre. Se ve claramente que los establecimientos para-hoteleros (en color celeste) tienen mayor ocupación que los hoteleros sólo en los meses de verano, lo cual resulta esperado considerando que los establecimientos para-hoteleros los conforman los hoteles sindicales, albergues, cabañas, bungalows, etc.

Figura 5. Tasa de Ocupación de Habitaciones por mes para la ciudad de Mar del Plata, 2005. Celeste: establecimientos para-hoteleros, Azul: establecimientos hoteleros



En la Figura 6 se muestra la relación entre tasa de Ocupación de Plazas y Personal por Plazas Ocupadas para establecimientos hoteleros y para-hoteleros. De acuerdo a lo esperado, la cantidad de personal por plaza ocupada es mayor para los establecimientos hoteleros.

Figura 6. Tasa de Ocupación de Habitaciones por mes para la ciudad de Mar del Plata, 2005. Anaranjado: establecimientos para-hoteleros, Rojo: establecimientos hoteleros



El programa utilizado en R para la construcción de la Figura 6 es el siguiente:

```
library(KernSmooth)
ancho1H<-dpill(MDP05H$top,MDP05H$ppo)
ancho1PH<-dpill(MDP05PH$top,MDP05PH$ppo)
plot(MDP05$top,MDP05$ppo,xlab='TOP',ylab='PPO',pch=20)
fit1H<- locpoly(MDP05H$top,MDP05H$ppo,bandwidth=ancho1H)
fit1PH<- locpoly(MDP05PH$top,MDP05PH$ppo,bandwidth=ancho1PH)
lines(fit1PH,col="brown",lty=2,lwd=2)
lines(fit1H,col="gold",lty=4,lwd=2)
```

4. DISCUSIÓN

En el presente trabajo se muestra la utilización del paquete Kernsmooth de R para la estimación de curvas de regresión polinómicas locales y para la construcción de bandas de variabilidad. Para ello se define en cada caso el grado del polinómio, el ancho de ventana y la función de núcleo (o de ponderación) deseados. Los programas son de construcción sencilla aún para aquellos usuarios poco familiarizados con la plataforma R. Si bien el paquete Kernsmooth sólo permite ajustes con anchos de ventana prefijados por el investigador (bandwidth), el usuario puede optar por otros paquetes de R (por ej. el paquete Stats) donde pueden construirse regresiones polinómicas locales usando anchos de ventana definidos como porcentaje de los datos totales que son vecinos al punto a estimar (span). Los suavizados de curvas mediante regresiones polinómicas locales resultan de gran utilidad para los analistas ya que constituyen una herramienta gráfica de sencilla interpretación y que son versátiles para "captar" los movimientos subyacentes en los datos. En la aplicación presentada, utilizando datos de la Encuesta de Ocupación Hotelera, se manifiestan estas características al poder explicar cada uno de los movimientos observados en las curvas, así por ejemplo se encuentran patrones claros sobre las tasas de ocupación de habitaciones a través del tiempo, comparaciones entre grupos de establecimientos hoteleros y para-hoteleros, etc.

A partir de este trabajo se planean futuras lineas de investigación en la profundización de los procedimientos de inferencia estadística de curvas de regresión por polinómios locales y la posibilidad de su aplicación en distintos software tanto comerciales como de distribución libre.

REFERENCIAS BIBLIOGRÁFICAS

Browman, A.W., Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-plus illustrations*. Oxford University Press.

Cleveland W. (1979) *Robust Locally Weighted Regression and Smoothing Scatterplots*. Journal of the American Statistical Association. Vol 74 pp 829-836

Fox, J. 2000. *Nonparametric Simple Regression: Smoothing Scatterplots*. Sage University Paper

Fox, J. 2000. *Multiple and Generalized Nonparametric Regression*. Sage University Paper

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Simonoff, J. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Venables, W.N., Ripley, B.D. (1999). *Modern Applied Statistics with S-plus*. New York: Springer-Verlag.

Wand,M.P.*The KernSmooth Package*.

http://cran.r-project.org/doc/packages/KernSmooth.pdf

Wand, M.P.; Jones, M.C. (1995) *Kernel Smoothing.* Chapman & Hall. London

FUENTE

Encuesta de Ocupación Hotelera (EOH), 2005. Instituto Nacional de Estadística y Censos (INDEC).