X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

APLICACIÓN PARA DETECTAR ATÍPICOS FUNCIONALES EN PROCESOS PRODUCTIVOS

Rodríguez, Mónica[1], Fernández, Rubén[1] y Vilar, Juan[1]

[1]Departamento de Matemáticas. Universidade da Coruña.

RESUMEN

En este trabajo se presenta una aplicación en el entorno estadístico R para detectar datos atípicos en un conjunto de datos funcionales. La motivación viene por el desarrollo de un proyecto realizado en colaboración con la empresa Ferrosolar S.L. y está pensada para analizar datos funcionales asociados a variables observadas en procesos productivos. El análisis estadístico realizado consta de dos etapas: en la primera se realiza un proceso de suavización y adecuación de los datos originales, en la segunda se obtienen gráficos y estadísticas descriptivas a partir de medidas de profundidad que permiten detectar datos atípicos en la muestra de datos funcionales.

Palabras y frases clave: Datos funcionales, valores atípicos, estimación no paramétrica, R, entorno gráfico.

1. INTRODUCCIÓN

El objetivo de muchos procesos industriales es la obtención de un material de interés a partir de una o varias materias primas. En muchos casos la etapa principal del proceso se realiza en un horno industrial y es posible obtener información de variables asociadas al proceso, algunas de estas variables pueden estar controladas por el usuario y otras no, siendo éstas últimas resultado del proceso. principalmente variables relacionadas con parámetros eléctricos del horno. La medición de una de estas variables en el tiempo que dura el proceso industrial es un dato funcional. Al repetir el proceso un número de veces, n, se dispone de una muestra de datos funcionales de la variable de interés y el objetivo del trabajo es el desarrollo de una aplicación capaz de detectar datos atípicos en esta muestra. Se ilustra el trabajo estudiando un conjunto de datos reales proporcionados por la empresa Ferrosolar S.L.

En muchos casos los datos de la variable de interés presentan irregularidades (ver Figura 1a). Cambios en otras variables controladas del proceso producen saltos en la respuesta. Además es habitual la presencia de atípicos e incluso de errores sistemáticos

de medición en determinados periodos, denominados saltos por contaminación. Por todo ello es necesario un tratamiento previo de las curvas originales para obtener la muestra de datos funcionales.

Este tratamiento consiste en un proceso de homogeneización y suavización de las curvas observadas. Se parte de una estimación aditiva, considerando los posibles factores que influyen en la respuesta y una componente aditiva no paramétrica función del tiempo. Así se obtienen estimaciones iniciales de los distintos componentes y residuos parciales eliminando el efecto de los factores. Estos residuos se utilizan para la corrección de los saltos por contaminación y la detección de atípicos, empleando la estimación de la varianza condicionada obtenida al aplicar regresión lineal local ponderada proporcionalmente a los residuos al cuadrado. Utilizando estos datos "procesados" y sin atípicos, se realiza una estimación aditiva final, siendo la componente no paramétrica un estimador lineal local con ventana corregida en base a la dependencia observada. La curva asociada a cada proceso es la estimación final del efecto parcial del tiempo (ver Figura 1b).

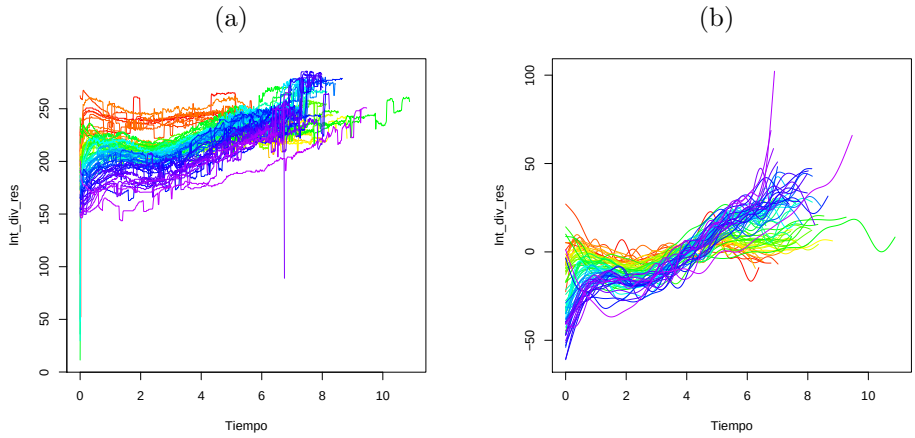

Figura 1: Ejemplo de datos originales (a) y curvas obtenidas después del proceso de suavizado (b).

Además, como la duración del proceso productivo es variable, en el problema en estudio, las curvas tienen diferente longitud (ver Figura 1), lo que origina un inconveniente ya que la mayor parte de las técnicas estadísticas disponibles para el análisis de datos funcionales son para curvas de la misma longitud. Para solucionar este problema se consideraron dos alternativas: una, se basa en truncar las curvas en un determinado punto, por ejemplo, el de la longitud de la curva más corta (ver Figura 2a); la segunda consiste en reescalar todas las curvas a un soporte común (ver Figura 2b).

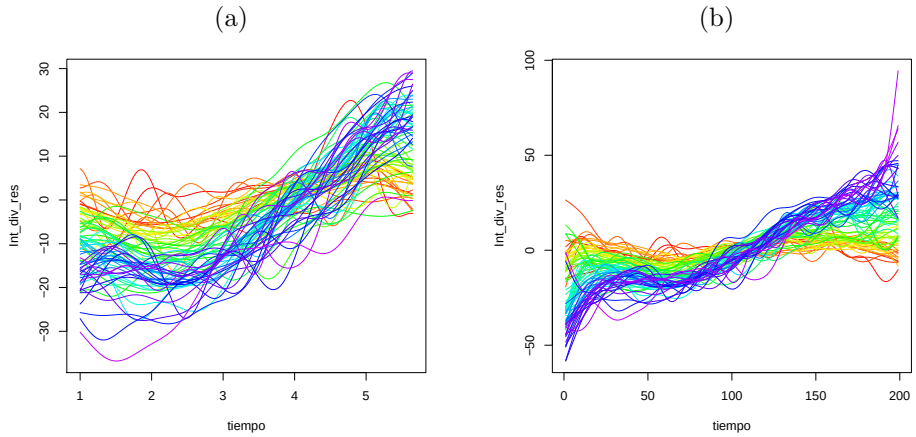

Figura 2: Conjuntos de datos funcionales obtenidos truncando (a) y reescalando (b) las curvas suavizadas de la variable de interés.

La organización del trabajo es la siguiente: en la sección 2 se presentan los métodos estudiados para la detección de atípicos funcionales, en la sección 3 se muestran algunos detalles sobre la aplicación gráfica desarrollada y finalmente en la sección 4 se exponen algunas conclusiones.

2. MÉTODOS PARA LA DETECCIÓN DE ATÍPICOS FUNCIONALES

La identificación de datos funcionales atípicos puede presentar bastantes dificultades. En algunos casos, las curvas pueden estar alejadas del resto de las curvas (atípicos por magnitud), en otros, pueden estar dentro del rango de las otras curvas pero con forma muy diferente (atípicos por forma) o pueden presentar una combinación de ambos casos. La mayoría de las técnicas disponibles para detectar atípicos funcionales se basan en una medida de profundidad. Estas medidas cuantifican la centralidad de una curva dentro de un conjunto de ellas. Curvas centrales se corresponden con valores grandes de profundidad (se define la mediana como la curva con mayor profundidad) y los datos atípicos se corresponden con valores "significativamente" pequeños de profundidad.

Como punto de partida se consideró el trabajo de Hyndman y Shang (2010) que utiliza dos medidas de profundidad basadas en las dos primeras componentes principales (robustas) de las curvas, transformando el problema funcional en un problema bidimensional. A partir de la muestra de datos bidimensionales de las dos primeras componentes principales se calcula la profundidad de Tukey (Tukey, 1975) o, alternativamente, se emplea una estimación no paramétrica de la densidad bidimensional como medida de profundidad, que se denota HDR, "highest density regions".

Para el análisis visual de los datos se emplearon gráficos rainbow y boxplots funcionales. En el gráfico rainbow se representan las curvas con diferentes colores de-

pendiendo de un orden, que puede ser el número de colada (Figura 2) o por la profundidad asociada la curva (Figura 5). Los colores que se asocian a cada curva para su representación son los colores del arco iris, de forma que una mayor profundidad se corresponde con el rojo (en negro se representa la mediana) y una menor con el violeta.

El boxplot funcional se contruye de modo análogo al caso de datos puntuales. Se representa una región interior que contiene el 50% de las curvas centrales, la mediana y una región exterior que contiene las observaciones no consideradas atípicas. Para determinar los valores atípicos habitualmente se sigue el procedimiento tradicional de inflar la región central (1.5 veces p.e.) o considerar atípico el dato con probabilidad estimada inferior al 1%. Por ejemplo, en la Figura 3, se muestra el denominado *bagplot funcional* de Hyndman y Shang (2010), obtenido al aplicar el bagplot bidimensional de Rousseeuw et al. (1999) a las puntuaciones de las dos primeras componentes principales, considerando la profundidad de Tukey e inflando 2.58 la region central en la escala de las componentes principales.

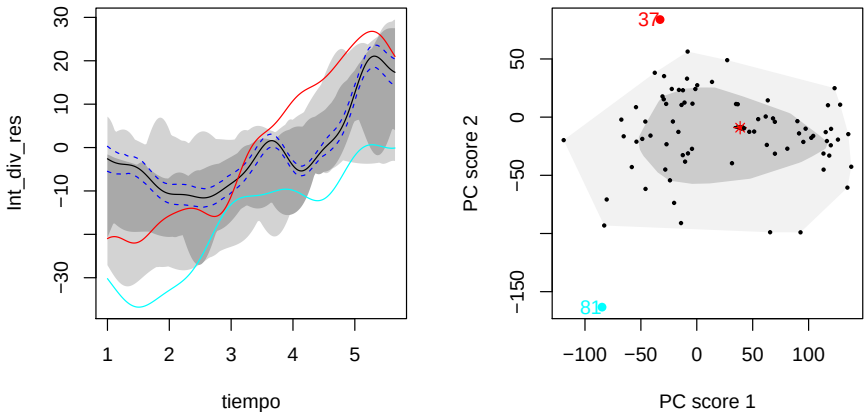

Figura 3: Boxplot funcional obtenido con la distancia de Tukey y correspondiente bagplot bidimensional de las dos primeras componentes principales.

3. DESARROLLO DE LA APLICACIÓN

La aplicación ha sido desarrollada bajo el entorno estadístico R (`http://www.r-project.org/`). R es un potente lenguaje funcional de software libre que puede extenderse fácilmente mediante paquetes. En el desarrollo de la interfaz gráfica se utilizó la librería GTK+ (`http://www.gtk.org/`), empleando *Glade* (`http://glade.gnome.org/`) para el diseño de la ventana principal. Se utilizaron los paquetes *RGtk2* y *RGtk2Extras* para gestionar los distintos objetos (widgets) de la interfaz gráfica desde R. En el suavizado de los datos se emplearon los paquetes *mgcv* y *KernSmooth*. Finalmente, se ha utilizado el paquete *rainbow* para el tratamiento de los datos funcionales,

aunque fué necesario la modificacion del código para mejorar su funcionalidad, corregir errores y reducir el tiempo de CPU.

La ventana principal de la aplicación consta de dos zonas diferenciadas (ver Figura 4). En la parte superior se encuentran las principales opciones del análisis. El menú archivo permite cargar un conjunto de datos y, posteriormente, el usuario debe seleccionar en la barra de tareas la variable con la que quiere trabajar. También debe elegir entre truncar o reescalar las curvas y la profundidad a utilizar (actualmente están disponibles la distancia de Tukey y la HDR, descritas en la sección 2). En la parte inferior de la ventana se muestran los resultados obtenidos. Pulsando en las distintas pestañas se puede alternar entre gráficos rainbow (ordenados por número de colada o valor de profundidad), boxplots o resultados numéricos (Figura 5).

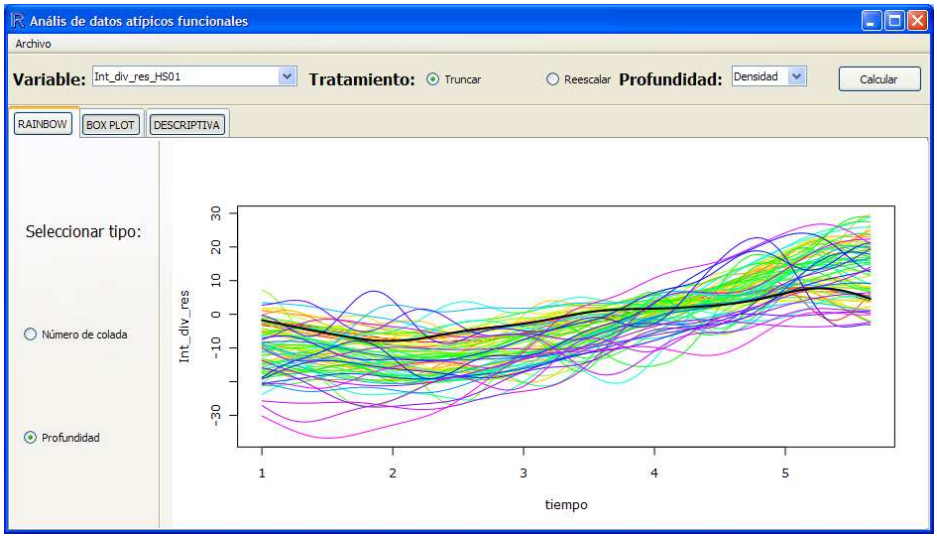

Figura 4: Ventana principal de la aplicación con gráfico rainbow coloreado según la profundidad de Tukey.

4. CONCLUSIONES

El empleo del entorno estadístico R con el paquete RGtk2 facilita el desarrollo de entornos gráficos para la aplicación de técnicas estadísticas avanzadas. Este tipo de herramientas resultan de gran utilidad porque permiten realizar análisis avanzados a usuarios no expertos o no familiarizados con el entorno R. La aplicación desarrollada responde a unos objetivos e intereses concretos de la empresa colaboradora, pero puede ser adaptada fácilmente a otros casos o incluso emplearse como punto de partida en el desarrollo de una aplicación de caracter más general para el análisis de datos funcionales.

Entre los trabajos pendientes está la implementación de nuevas medidas de profundidad (por ejemplo, las basadas en proyecciones aleatorias, ver Febrero et al. (2007 y 2008)) y el estudio de su adecuación a este problema concreto.

Figura 5: Ventana principal con descriptivos de los valores de profundidad y listado de datos funcionales con profundidad asociada.

Agradecimientos: Los autores desean agradecer la colaboración prestada por la empresa Ferrosolar S.L. Este trabajo ha sido financiado parcialmente por los proyectos MEC MTM2008-00166 y MTM2008-03010.

REFERENCIAS

Febrero, M., Galeano, P. y Gonzalez-Manteiga, W. (2007) A Functional Analysis of NOx Levels: Location and Scale Estimation and Outlier Detection. Computational Statistics, 22 , 411–427.

Febrero, M., Galeano, P. y Gonzalez-Manteiga, W. (2008) Outlier Detection in Functional Data by Depth Measures, With Application to Identify Abnormal NOx Levels. Environmetrics, 19, 331–345.

Hyndman, R. y Shang H. (2010) Rainbow Plots, Bagplots, and Boxplots for Functional Data. Journal of Computational and Graphical Statistics, 19, 29-45.

Rousseeuw, P., Ruts, I. y Tukey, J. (1999) The Bagplot: A Bivariate Boxplot. The American Statistician, 53 , 382–387.

Tukey, J. (1975) Mathematics and the Picturing of Data. Proceedings of the International Congress of Mathematicians, Vol. 2, ed. R. D. James, pp. 523–531.