Informe Práctica Profesional

Método Hidiroglou-Berthelot y Distancias Intercuartiles para la detección de puntos atípicos en datos de encuestas económicas

Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Juan Pablo Zamora Alarcón

Profesor Guía: Mauricio Castro

15 de diciembre del 2021, Santiago de Chile.

Índice

1 Introducción

En el siguiente documento presentamos y evaluamos dos métodos para la detección de puntos atípicos o outliers. Además, para la evaluación de los datos se analiza el ratio intermensual, definido a continuación.

Sea $t \in \mathbb{N}$, el periodo, y los pares de datos $(x_i(t), x_i(t+1))$, con $i = \{1, ..., n\}$ tal que:

$$r_i = x_{i,t}/x_{i,t-1}$$

El ratio intermensual, lo aplicaremos para calcular las bandas de aceptación en el método de distancias intercuartiles (Hunt,et al.,1999,p.540) y en el método de Hidiroglou-Berthelot (Hidiroglou,et al.,1986,p.77).

Primero, el método de las distancias intercuartiles, identifica los valores outliers como todos aquellos datos que no pertenecen al intervalo:

$$(Q_1 - k \cdot IQR, Q_3 + k \cdot IQR)$$

Donde Q_1 es el primer cuartil, Q_3 es el tercer cuartil, IQR es el rango intercuartil definido como $IQR := (Q_3 - Q_1)$, y k es una constante. Para efectos del presente informe, tomaremos $k = 1.5$ (D'Orazio,2017,p.8).

Segundo, el método de Hidiroglou-Berthelot, identifica los valores outliers como todos aquellos datos que no pertenecen al intervalo:

$$(E_M - C \cdot dq_1, E_M + C \cdot dq_3)$$

En este método se definen dos transformaciones S_i y E_i.

La primera transformación es de los ratios intermensual r_i a S_i, donde hay dos casos, $S_i := 1 - \frac{r_M}{r_i}$, si $0 < r_i < r_M$ y $S_i := \frac{r_i}{r_M} - 1$, si $r_i \geq r_M$, con r_M mediana de los ratios. La segunda transformación es E_i con $E_i := S_i \cdot max(x_i(t), x_i(t+1))^U$. Dado lo anterior, se definen dos desviaciones, la primera dq_1, como $dq_1 := max(E_M - E_{q_1}, |A \cdot E_M|)$, y la segunda desviación dq_3, como $dq_3 := max(E_{q_3} - E_M, |A \cdot E_M|)$, definidas como la banda inferior y la banda superior respectivamente. Importante especificar que E_M es la transformación de la mediana de los datos, E_{q_1} es la transformación del primer cuartil de los datos y, E_{q_3} es la transformación del tercer cuartil de los datos. Por último, observar que $|A \cdot E_M|$ evita los falsos outliers, cuando $E_M - E_{q_1}$ y $E_{q_3} - E_M$ son pequeños. El parámetro C controla principalmente el ancho de la banda de aceptación, y el parámetro U, si es un valor alto, otorga más importancia a magnitudes altas, en cambio, si U es un valor pequeño, otorga más importancia a magnitudes bajas.

Dado lo anterior, observamos que la elección de k, en el método de distancias intercuartiles, y los parámetros A, C y U en el método de Hidiroglou-Berthelot, es subjetiva y puede ocacionar la no detección de outliers, como también, fenómenos donde un outliers es no detectado porque hay presencia de otros outliers muy próximos entre si ("Making Effect") (Ishikawa,et al.,2010,p.7) y, otro caso descrito es cuando una buena observación es identificada incorrectamente como un outliers, ya que, hay presencia de un subconjunto de datos sin outliers ("Swamping Effect") (Chiang,2007,p.301). Dado lo anterior, detallaremos la construcción de una grilla de datos para el método de Hidiroglou-Berthelot (Hunt,et al.,1999,p.541), por convención, se mantiene constante el valor $A = 0.05$ (Hunt,et al.,1999,p.541). Estos valores se entrenan y el resultado obtenido, será aquel que obtenga el menor porcentaje promedio de outliers, para cada parámetro.

Por último, se aplicarán estos dos métodos a una misma encuestas económica. Donde observamos las cantidades detectadas de outliers, al evaluar con distintos valores los parámetros de ambos métodos antes mencionado.

2 Datos Encuesta Económica.

2.1 Base de datos.

A continuación, se presenta un ejemplo de salida de la base de datos, desde enero hasta diciembre del 2014, para la primera empresa.

```
##   VTA_ENERO14 VTA_FEBRERO14 VTA_MARZO14 VTA_ABRIL14 VTA_MAYO14 VTA_JUNIO14
## 1   327724167     252090106   260309432   274450681  284953957   242126857
##   VTA_JULIO14 VTA_AGOSTO14 VTA_SEPTIEMBRE14 VTA_OCTUBRE14 VTA_NOVIEMBRE14
## 1   272594131    287348279        265172147     302804081        93391844
##   VTA_DICIEMBRE14
## 1       310069569
```

La dimensión de la base es de 71 Empresas (observaciones) y registro de ventas de 78 meses (parámetros), desde enero del 2014 hasta junio del 2020.

2.2 Código Venta intermensual.

Primero, presentamos el código para el ratio de venta intermensual (r_1). Incorporando ejemplos de cada salida.

```
r_1 = vector(mode = "list",n) # ratio de venta intermensual de cada empresa
for( j in 1:n){ # n = n° de observaciones
  for(i in 1:(p-1)) # p = n° de parámetros
    r_1[[j]][i] = x[j, i + 1]/x[j, i]
}
```

```
## [1] "Ratio de venta intermensual Febrero/Enero 2014 para la empresa N°1 :0.769"
```

```
## [1] "Ratio de venta intermensual Marzo/Febrero 2014 para la empresa N°2 :0.988"
```

3 Método de Distancias Intercuartiles de los ratios históricos.

3.1 Código Rango Intercuartíl de ratio intermensual.

A continuación, comenzaremos trabajando con el código del rango intercuartíl ratio intermensual para k = 1.5.

```
lowlimit_1.5 = numeric(length = n) # guardamos el límite inferior
suplimit_1.5 = numeric(length = n) # guardamos el límite superior
minimo = numeric(length = n) # guardamos el valor mínimo detectado
maximo = numeric(length = n) # guardamos el valor máximo detectado
r1_25 = numeric(length = n) # guardamos el Q1 de los ratios
r1_M = numeric(length = n) # guardamos el Q2 de los ratios
r1_75 = numeric(length = n) # guardamos Q3 de los ratios
B_1.5 = matrix(NA, ncol = 3, nrow = n)
for( i in 1:n){
    minimo[i] = min(r_1[[i]], na.rm = TRUE)
```

```
    maximo[i] = max(r_1[[i]], na.rm = TRUE)
    r1_25[i] = as.numeric(quantile(r_1[[i]], 1/4, na.rm = TRUE))
    r1_M[i] = as.numeric(quantile(r_1[[i]], 1/2, na.rm = TRUE))
    r1_75[i] = as.numeric(quantile(r_1[[i]], 3/4, na.rm = TRUE))
    lowlimit_1.5[i] = max(minimo[i], r1_M[i] - 1.5*(r1_M[i] - r1_25[i]), na.rm = TRUE)
    suplimit_1.5[i] = min(maximo[i],r1_M[i] + 1.5*(r1_75[i] - r1_M[i]), na.rm  = TRUE)
    B_1.5[i,2]  = suplimit_1.5[i] # matriz con las bandas de confianza superior
    B_1.5[i,3]  = lowlimit_1.5[i] # matriz con las bandas de confianza inferior
    B_1.5[i,1] = i
}
```

3.2 Detección de outliers para ratio intermensual.

En el siguiente código, se realiza la detección de outliers con k = 1.5.

```
sup_out = vector( mode = "list",n) # guardamos los outliers detectados
inf_out= vector(mode = "list",n) # guardamos los outliers detectados
for( j in 1:n ){
  for( i in 1:(p-1)){
    if(is.na(r_1[[j]][i]) == TRUE){
    }else{
      if(as.numeric(r_1[[j]][i]) > B_1.5[j,2]){
        # guardamos los outliers detectados
        sup_out[[j]][i] = as.numeric(r_1[[j]][i])
      }else{
        if(as.numeric(r_1[[j]][i]) <= B_1.5[j,2]){
          sup_out[[j]][i]  = 0 # serán 0 los no outliers
        }
      }
      if(as.numeric(r_1[[j]][i]) < B_1.5[j,3]){
        # guardamos los outliers detectados
        inf_out[[j]][i]  = as.numeric(r_1[[j]][i])
      }else{
        if(as.numeric(r_1[[j]][i]) >= B_1.5[j,3]){
          inf_out[[j]][i]  = 0 # serán 0 los no outliers
        }
      }
    }
  }
}
```

3.3 Resumen de outliers por empresa y porcentaje detectado.

Teniendo presente lo anterior, ahora mostraremos salidas computacionales, con el resumen del porcentaje de outliers detectados de las primeras diez empresas.

Porcentaje Promedio de Outliers ratio intermensual con k = 1.5
31.5

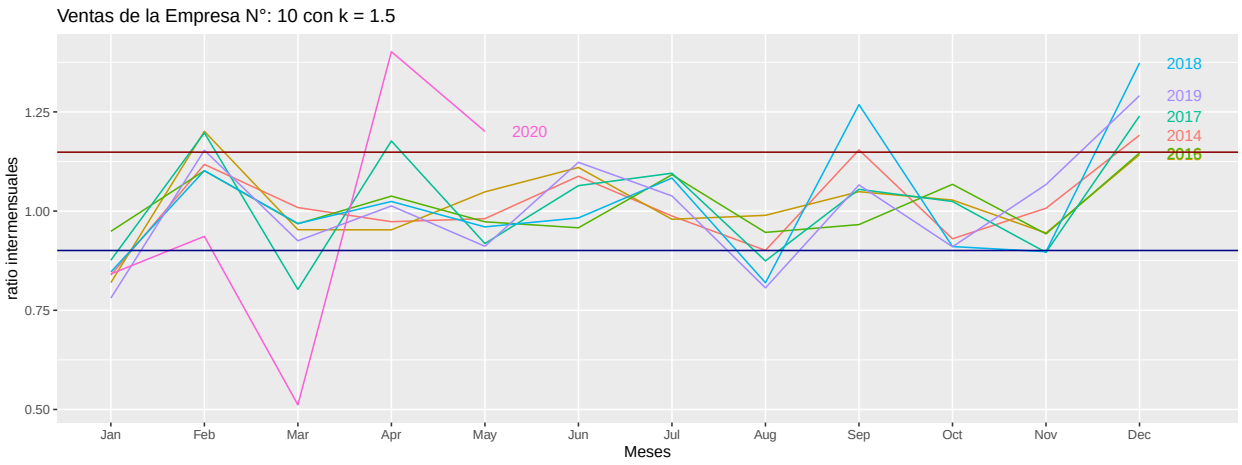
EMPRESA	Total de datos	Mínimo	Q1	Mediana	Q3	Máximo
1	77	0.0000000	0.9166788	0.9915638	1.063276	3.320093
2	77	0.1229795	0.8956638	0.9805354	1.079085	11.288468
3	77	0.3838195	0.9167730	1.0000000	1.088255	7.265570
4	77	0.0687373	0.9425828	1.0338920	1.096113	9.198854
5	77	0.1720810	0.9170997	0.9975688	1.044937	5.665234
6	77	0.4642076	0.9392346	1.0057779	1.073659	1.395634
7	77	0.3156325	0.8868969	1.0079402	1.187531	4.064854
8	77	0.1014644	0.9246484	1.0111955	1.103657	9.289229
9	77	0.2867899	0.8907405	1.0005246	1.074870	3.576568
10	77	0.5116542	0.9302196	0.9893028	1.095454	1.401752

EMPRESA	Banda Inferior	Banda Superior	Total de Outliers	Outliers %
1	0.8792363	1.099133	30	38.96
2	0.8532280	1.128359	23	29.87
3	0.8751595	1.132382	30	38.96
4	0.8969281	1.127224	29	37.66
5	0.8768652	1.068620	27	35.06
6	0.9059629	1.107600	27	35.06
7	0.8263753	1.277326	29	37.66
8	0.8813748	1.149888	23	29.87
9	0.8358485	1.112042	22	28.57
10	0.9006781	1.148530	25	32.47

3.4 Gráfico con identificación de outliers.

Ahora, presentamos un ejemplo de gráfico de ventas, para ratios intermensual k = 1.5.

Lo anterior, para la empresa n° 10, identificando las bandas de confianza respectivamente.



4 Método Hidiroglou-Berthelot.

4.1 Función S para el ratio intermensual.

A continuación definimos la transformación S para el ratio intermensual.

```
s1 = vector(mode = "list",n) # guardamos los valores de la transformación S_i
for( j in 1:n ){
  # la mediana de los ratios calculada por empresa
```

```
  r_M = as.numeric(quantile(r_1[[j]], 1/2, na.rm = TRUE))
  for( i in 1:length(r_1[[j]])){
    if(is.na(r_1[[j]][i]) == TRUE){
    }else{
      if( 0 < as.numeric(r_1[[j]][i]) & as.numeric(r_1[[j]][i]) < r_M){
        s1[[j]][i] = (1 - (r_M/as.numeric(r_1[[j]][i])))
      }else{
        s1[[j]][i] = ((as.numeric(r_1[[j]][i])/r_M) - 1)
      }
    }
  }
}
```

4.2 Grilla de los parámetros A, U y C.

Se presenta una grilla de valores, donde A = 0.05 es fijo, C toma valores desde 5 a 75, cada 5 unidades. Y el parámetro U, toma valores desde 0 a 1, cada 0.1 unidades. Se muestran una salida computacional de ejemplo.

A	U	C
0.05	0.0	5
0.05	0.1	5
0.05	0.2	5
0.05	0.3	5
0.05	0.4	5
0.05	0.5	5
0.05	0.6	5
0.05	0.7	5
0.05	0.8	5
0.05	0.9	5
0.05	1.0	5

4.3 Código de la tranformación E, las desviaciones e identificación de outliers ratio intermensual.

A continuación, incorporando la grilla de valores de los parámetros A, C y U, calcularemos la transfomación E, las desviaciones e identificando los outliers para cada empresa.

Primero, calculamos la transformación E y las desviaciones dq_1 y dq_3. Segundo, identificamos los valores que son outliers, es decir aquellos valores que no pertenecen al intervalo $(E_M - Cdq_1, E_M + Cdq_3)$, con E_M es la transformación de la mediana para cada empresa.

Por último, se presenta una salida computacional, con los valores de A, C y U, con los cuales se obtiene el menor porcentaje promedio de outliers identificados.

```
# creación de la grilla de valores.
tunegrid1 = expand.grid(A = c(0.05),U = seq(0,1, 0.1),C = seq(5,75,5))
error1=rep(NA,nrow(tunegrid1))
for(t in 1:nrow(tunegrid1)){
  E = vector(mode = "list",n) # guardamos los valores de la transformación E
  # guardamos los valores de la transformación del Q1
  E_25 = vector(mode = "list",n)
  # guardamos los valores de la transformación del Q2
  E_M = vector(mode = "list",n)
  # guardamos los valores de la transformzacion del Q3
  E_75 = vector(mode = "list",n)
```

```
  dq1 = vector(mode = "list",n) # guardamos los valores de la desviación de Q1
  dq3 = vector(mode = "list",n) # guardamos los valores de la desviación de Q3
  sup_out = vector( mode = "list",n) # guardamos los outliers detectados
  inf_out = vector( mode = "list",n) # guardamos los outliers detectados
  sup = numeric(length = n) # guardamos la cantidad de outliers detectados
  inf = numeric(length = n) # guardamos la cantidad de outliers detectados
  sum = 0
  for( j in 1:n){
    for(i in 1:(p-1)){
      if(is.na(x[j,i]) == TRUE | is.na(x[j,i + 1]) == TRUE){
      }else{
        E[[j]][i] = s1[[j]][i]*(max(x[j,i],x[j, i + 1 ]))^tunegrid1$U[t]
        E_25[[j]] = as.numeric(quantile(E[[j]], 1/4, na.rm = TRUE))
        E_M[[j]] = as.numeric(quantile(E[[j]], 1/2, na.rm = TRUE))
        E_75[[j]] = as.numeric(quantile(E[[j]], 3/4, na.rm = TRUE))
        dq1[[j]] = max(E_M[[j]] - E_25[[j]], abs(tunegrid1$A[t]*E_M[[j]]))
        dq3[[j]] = max(E_75[[j]] - E_M[[j]], abs(tunegrid1$A[t]*E_M[[j]]))
      }
    }

    for( i in 1:length(E[[j]])){
      if(is.na(E[[j]][i]) == TRUE | is.na(E_M[[j]]) == TRUE |
         is.na(dq3[[j]]) == TRUE | is.na(dq1[[j]]) == TRUE ){
      }else{
        if( E[[j]][i] >  (E_M[[j]] + tunegrid1$C[t]*dq3[[j]])){
          sup_out[[j]][i] = E[[j]][i]
        }else{
          if( E[[j]][i] <= (E_M[[j]] + tunegrid1$C[t]*dq3[[j]])){
            sup_out[[j]][i] = "NOT" # Serán NOT los valores no outliers
          }
        }
        if( E[[j]][i] < (E_M[[j]] - tunegrid1$C[t]*dq1[[j]])){
          inf_out[[j]][i] = E[[j]][i]
        }else{
          if(E[[j]][i] >= (E_M[[j]] - tunegrid1$C[t]*dq1[[j]])){
            inf_out[[j]][i] = "NOT" # Serán NOT los valores no outliers
          }
        }
      }
    }
    sup[j] = length(which(sup_out[[j]] != "NOT"))
    inf[j] = length(which(inf_out[[j]] != "NOT"))
    sum = round(((sup[j] + inf[j])*100)/length(E[[j]]),2) + sum
  }
  # calcula el % de outliers promedio para cada empresa
  error1[t] = round(sum/n,2)
}
# Identifica el valor del parámetro U con el menor error.
best.U = tunegrid1$U[which.min(error1)]
# Identifica el valor del parámetro C con el menor error.
best.C = tunegrid1$C[which.min(error1)]
# Identifica el valor del parámetro A con el menor error.
best.A = tunegrid1$A[which.min(error1)]
best1 = cbind(best.U,best.C,best.A,min(error1))
```

```
best1 = data.frame(best1)
colnames(best1) = c("U","C","A", "Outliers identificados intermensual %:")
```

4.4 Evaluando los valores encontrados de los parámetros.

A continuación, evaluamos los valores encontrados en el mismo código, y presentamos tablas de resumen.

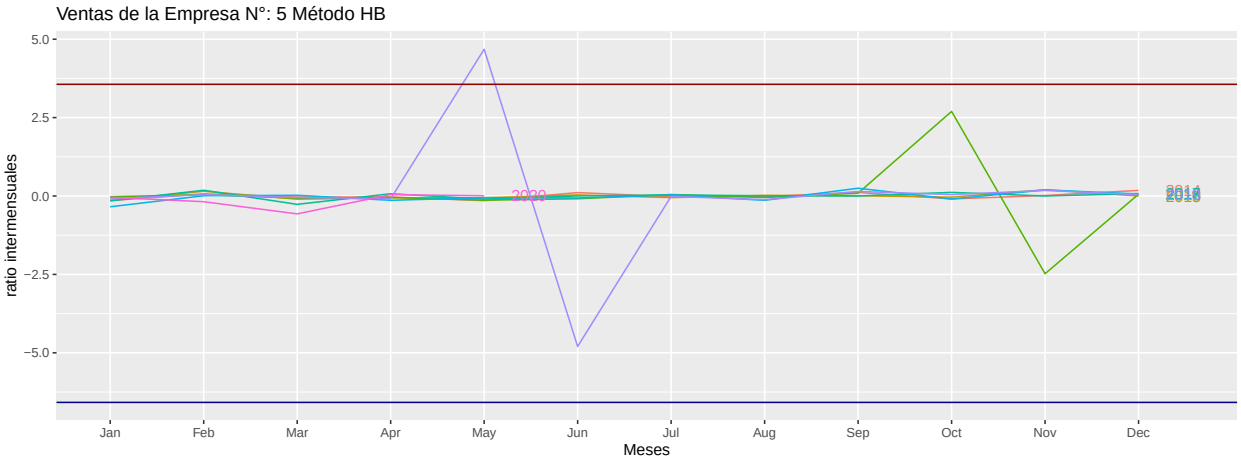
U	C	A	Porcentaje Promedio de Outliers ratio intermensual:
0	75	0.05	0.84

EMPRESA	Total de datos	Mínimo	Q1	Mediana	Q3	Máximo
1	77	-2.214944	-0.08169169	0	0.07232265	2.34834
2	77	-6.973159	-0.09475838	0	0.1005057	10.51255
3	77	-1.605391	-0.09078257	0	0.08825472	6.26557
4	77	-14.0412	-0.09687343	0	0.06018178	7.897307
5	77	-4.79709	-0.08774299	0	0.04748321	4.679041
6	77	-1.166656	-0.0708485	0	0.06749126	0.3876165
7	77	-2.193398	-0.1364796	0	0.1781756	3.032833
8	77	-8.966015	-0.0936	0	0.09143828	8.186383
9	73	-2.488702	-0.1232504	0	0.07430596	2.574693
10	77	-0.9335379	-0.06351526	0	0.1072991	0.4169093

EMPRESA	Banda Inferior	Banda Superior	Total de Outliers	Outliers %
1	-6.126877	5.424199	0	0.0
2	-7.106879	7.537925	1	1.3
3	-6.808693	6.619104	0	0.0
4	-7.265508	4.513634	2	2.6
5	-6.580724	3.561241	1	1.3
6	-5.313637	5.061845	0	0.0
7	-10.235969	13.363173	0	0.0
8	-7.020000	6.857871	2	2.6
9	-9.243777	5.572947	0	0.0
10	-4.763644	8.047434	0	0.0

4.5 Gráfico con identificación de outliers.

A continuación, presentamos un gráficos con las ventas de ratio intermensuales para la empresa N° 5. Cada gráfico tiene su respectiva banda de confianza.



5 Conclusión

En el trabajo presentado, se comparan dos métodos para la detección de datos atípicos. Para efectos de este informe se realiza el trabajo para una sola encuesta económica, y principalmente analizando el comportamiento de ventas de ratio intermensual.

Primero, diseñamos e implementamos un código para detectar la cantidad de outliers, el porcentaje promedio de outliers, las bandas de confianza, entre otros resultados a nivel transversal.

En el caso de distancias intercuartiles, se realiza la detección de datos atípicos para $k = 1.5$ (D'Orazio,2017,p.8). Donde el parámetro k, no tiene una grilla de valores incorporada, sin embargo, el código está diseñado, tal que, el valor de k puede ser modificado, por ejemplo, con $k = 2$ o $k = 3$ (D'Orazio,2017,p.8). Para este caso se detectó un porcentaje promedio de detección de outliers de 31.5%.

Para el método Hidiroglou-Berthelot, se utilizó una grilla de valores, tal que, tomamos $A = 0.05$ (Hunt,1999,et al.,p.541) constante, modificando el resto de los parámetros. Así, para la base trabajada tenemos como resultado $U = 0$, $C = 75$ y $A = 0.05$ con un porcentaje promedio de detección de outliers de 0.84%. Además de lo anterior, el código diseñado e implementado para la grilla se puede modificar, tal que, el porcentaje promedio de outliers detectados, evaluando todos los valores de la grilla, sea menor a un 5%. También, podemos modificar la grilla, tal que el parámetro $A = \{0.05, 0.2\}$ (Denmark,et al.,p.8). Importante a tomar en consideración, que el tiempo de ejecución (costo computacional) del código para detectar outliers, en casos de ratios interanuales e intermensuales, es alto (aproximadamente 40 minutos). Tal desventaja, no se pudo solucionar manteniendo la misma grilla de valores, detallada en los ítems anteriores.

Se incorporaron en cada ítem, un gráfico con sus respectivas bandas de confianza. Se optó, por comodidad por la empresa N°5 y así poder comparar con las tablas de resumen de cada sección, comprobando que el algoritmo funciona de manera correcta. Cabe mencionar, que no se especifica en las tablas de resumen, la cantidad de valores faltantes de cada empresa, sin embargo, cada base de dato sí contiene tiene valores faltantes, los cuales, se omiten tanto para los cálculos de las bandas de confianza, y los gráficos de ventas.

En conclusión, los resultados expuestos en cada ítem, muestran una detección de datos atípicos efectiva, para ambos métodos. Especificando la cantidad de outliers detectados, los cuartiles, los valores máximos, los valores mínimos, las bandas de confianza y los porcentajes de outliers por empresa.

Por último, se deja como sugerencia de estudio, la implementación de la metodología de aprendizaje no-supervisado K-Means, expuesta en (Evangelos Pongas, 2011), para la identificación de outliers en encuestas económicas.

6 Referencias

Hunt, J. W., Johnson, J. S., & King, C. S. (1999). Detecting outliers in the monthly retail trade survey using the Hidiroglou-Berthelot method. In Proceedings of the Section on Survey Research Methods of the American Statistical Association (pp. 539-543).

D'Orazio, M. (2017). Outlier Detection in R: Some Remarks. In 5th International Conference "New Challenges for Statistical Software–The Use of R in Official Statistics". Bucharest, Romania. http://www.r-project. ro/conference2017/presentations/D'Orazio-Outlier_Detection_in_R_ (slides_v5). pdf.

Chiang, J. T. (2007). The masking and swamping effects using the planted mean-shift outliers models. Int. J. Contemp. Math. Sciences, 2(7), 297-307.

Ishikawa, A., Endo, S., & Shiratori, T. (2010). Treatment of outliers in business surveys: The case of short-term economic survey of enterprises in Japan (Tankan). Bank of Japan, 2-1.

Hidiroglou, M. A., & Berthelot, J. M. (1986). Statistical editing and imputation for periodic business surveys. Survey methodology, 12(1), 73-83.

Denmark, S., & Larsen, M. B. Data editing in the Danish CPI.

Evangelos Pongas, OECD November 2011. Using cluster analysis for Identifying outliers and possibilities offered when calculating Unit Value Indices.