MÁSTER EN MATEMÁTICA COMPUTACIONAL

PROYECTO FINAL DE MÁSTER

MÉTODOS PARA EL ESTUDIO DE LA DEGRADACIÓN DE CONTADORES DE AGUA SOBRE DATOS DE CONSUMO

Autor
José Rodríguez Cuaresma

Tutores académicos
Pablo Gregori Huerta

Castellón, octubre de 2018

Resumen

En la gestión del agua, dentro del sistema de redes de abastecimiento, podemos realizar un balance hídrico propuesto por la International Water Association (IWA) partiendo del volumen de agua introducido al sistema.

Este volumen introducido en la red de abastecimiento se dividirá en un principio en dos grandes conjuntos de salida. El primero de ellos se corresponde a la salida mediante un consumo autorizado, el cual podremos medir y registrar. En segundo lugar, analizaremos unas pérdidas, subdivididas en pérdidas reales (como pueden ser fugas en la red, conexionados, etc.), y unas pérdidas aparentes o comerciales, debidas al consumo no autorizado y a errores de medición.

El apartado que nos ocupa en este proyecto está enmarcado dentro de estas pérdidas aparentes, concretamente en el error de medición, que es llevado a cabo por un dispositivo de medición al que se denomina “contador”.

En el año 2015 se finalizó un ambicioso proyecto dentro de la empresa FACSA para realizar un estudio sobre estas pérdidas aparentes. Para ello, se utilizaron una gran cantidad de contadores, que fueron analizados en bancos de prueba. Tras su análisis, se pudo obtener la evolución del error de medición de diferentes contadores, dependiendo de su edad y volumen total.

El objetivo principal de este proyecto es dar los primeros pasos para poder obtener unos resultados similares a los del estudio mencionado; partiremos de datos de consumo, de los que posteriormente obtendremos la evolución de este error, teniendo en cuenta más características del contador y su instalación.

Resum

En la gestió de l'aigua, dins del sistema de xarxes de proveïment, podem realitzar un balanç hídric proposat per la International Water Association (IWA) partint del volum d'aigua introduït al sistema.

Aquest volum introduït en la xarxa de proveïment es dividirà al principi en dos grans conjunts d'eixida. El primer d'ells es correspon a l'eixida mitjançant un consum autoritzat, el qual podrem mesurar i registrar. En segon lloc, analitzarem unes pèrdues, subdividides en pèrdues reals (com poden ser fugides en la xarxa, conexionats, etc.), i unes pèrdues aparents o comercials, degudes al consum no autoritzat i a errors de mesurament.

L'apartat que ens ocupa en aquest projecte està emmarcat dins d'aquestes pèrdues aparents, concretament en l'error de mesurament, que és dut a terme per un dispositiu de mesurament al que es denomina “comptador”.

L'any 2015 es va finalitzar un ambiciós projecte dins de l'empresa FACSA per a realitzar un estudi sobre aquestes pèrdues aparents. Per a açò, es van utilitzar una gran quantitat de comptadors, que van ser analitzats en bancs de prova. Després de la seua anàlisi, es va poder obtenir l'evolució de l'error de mesurament de diferents comptadors, depenent de la seua edat i volum total.

L'objectiu principal d'aquest projecte és donar els primers passos per a poder obtenir uns resultats similars als de l'estudi esmentat; partirem de dades de consum, de les quals posteriorment obtindrem l'evolució d'aquest error, tenint en compte més característiques del comptador i la seua instal·lació.

Índice general

Capítulo 1

Introducción

En este primer capítulo se explica el contexto y motivación del proyecto desarrollado y se presentarán los objetivos generales. También se introducirá formalmente un balance hídrico como punto de partida a la problemática en la degradación de contadores y presentaremos la empresa FACSA, la cual nos ha facilitado la labor para poder realizar este proyecto. Por último, se mencionará el software utilizado para la realización del mismo.

En el resto de capítulos realizaremos un análisis descriptivo de los datos y expondremos los métodos de análisis propuestos para el estudio de la degradación de los contadores partiendo de datos de consumo.

1.1 Contexto y motivación del proyecto

Una red de abastecimiento de agua potable es un sistema de obras de ingeniería concatenadas que permiten llevar hasta la vivienda de los habitantes de una ciudad, pueblo o área rural con población relativamente densa, el agua potable.

Dentro de una red de abastecimiento, tradicionalmente se han buscado las pérdidas de agua como pérdidas reales, es decir, fugas dentro del mismo sistema. El paso del tiempo y las nuevas tecnologías han permitido también centrar los esfuerzos en las pérdidas aparentes, sobre todo en las de medición.

Por el momento no hay ningún método para poder obtener errores de medición mediante el estudio de datos de consumo, aunque intentaremos arrojar algo de luz con este proyecto.

La motivación de basar este tipo de estudios en los datos de consumo es debida a que estos datos ya son conocidos y se obtienen fácilmente en la lectura de contadores, por lo cual no se necesitaría un trabajo extra.

1.2 Objetivos generales

Como se ha comentado en el apartado anterior, no hay hasta la fecha ningún método establecido para el estudio de la degradación de contadores basado en datos de consumo.

Para poder realizar algún método de estudio, se realiza un análisis de los datos, presente en el análisis descriptivo de este proyecto, como punto de partida.

Sabiendo las características de los datos de partida, se propone como objetivo principal exponer dos métodos para el estudio de la degradación de los contadores, así como su aplicación y la exposición de las conclusiones.

Por tanto, y de manera general, el objetivo principal de este trabajo es poder dar los primeros pasos en el estudio de las pérdidas por errores de medición basadas en datos de consumo y ofrecer herramientas y procedimientos para su causa.

1.3 Software utilizado

Para la realización de los diferentes análisis sobre los datos se ha utilizado lenguaje R, y para su aplicación el programa Rstudio [14]

RStudio es una interfaz que permite acceder de manera sencilla a toda la potencia de R, para utilizar RStudio se requiere haber instalado R previamente.

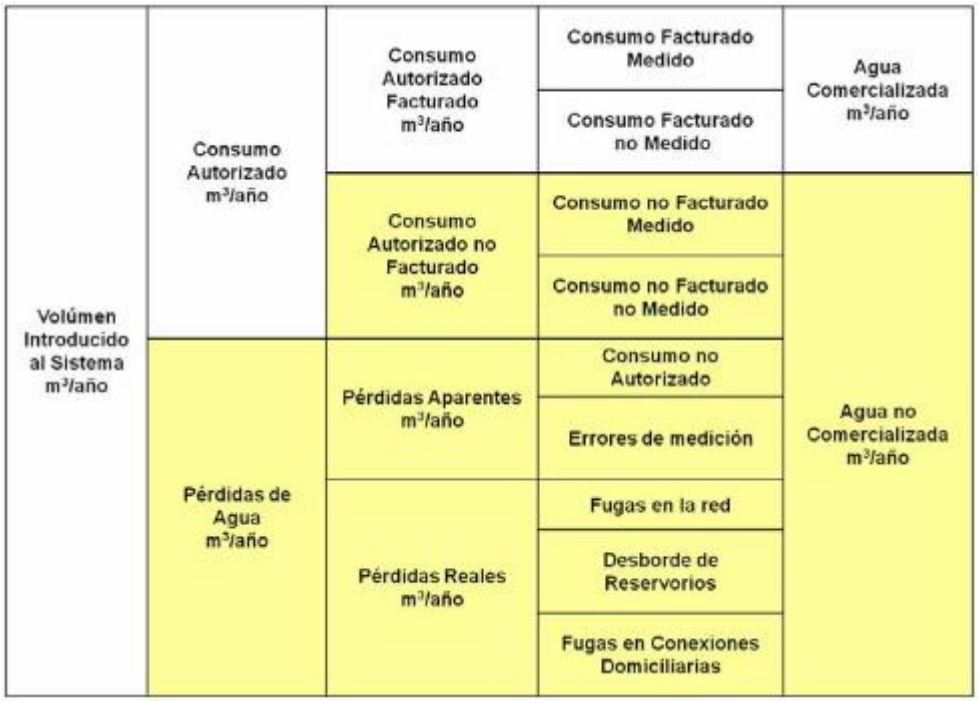
R es un lenguaje para el cálculo estadístico orientado a objetos y permite la generación de gráficos, que ofrece una gran variedad de técnicas estadísticas y descriptivas en un entorno de análisis y programación estadísticos.

Encontramos un lenguaje de programación completo con el que se añaden nuevas técnicas mediante la definición de funciones. Entre sus muchas ventajas, permite autocompletar, incluye un menú de ayuda muy completo, dispone de un depurador de código que detecta posibles errores, es multiplataforma (existen versiones para Windows, Linux y Mac…) y además es de libre distribución.

R se puede conseguir gratuitamente en varios sitios web. Una de ellos es http://www.r-project.org [1]

1.4 Balance hídrico: errores de medición

El balance hídrico dentro de una red de abastecimiento consiste a grandes rasgos en un balance entre las entradas y salidas de agua. Según la International Water Association (IWA) podemos diferenciar el tipo de salida de la red.



Resulta evidente que las pérdidas de agua podrían definirse como la diferencia entre el volumen de entrada al sistema y el consumo autorizado. [2]

Dentro de este balance, estamos focalizados en las pérdidas aparentes o comerciales y, dentro de las mismas, en las pérdidas por error de medición.

¿Por qué es tan importante minimizar estas pérdidas?

Podemos dar varias respuestas a esta pregunta. A todos, en un primer momento, se nos viene a la mente el aspecto económico, ya que el agua no contabilizada es agua no facturada, pero dejándolo a un lado, caben destacar también otros aspectos. El primero de ellos es que el agua es un bien común que debe ser utilizado de la manera más eficiente posible. El segundo señala que grandes pérdidas indican una gestión ineficiente por parte de la empresa gestora, y por tanto, un mal funcionamiento de la red de abastecimiento que afecta a los usuarios de la misma.

1.5 FACSA

1.5.1 La empresa

FACSA es la empresa privada española con más experiencia en la gestión del ciclo integral del agua.

Perteneciente al Grupo Gimeno, se fundó en Castellón en el año 1873 con un primer objetivo: dotar a la capital de la provincia de una moderna red de distribución de agua potable. Desde entonces, ha diversificado sus actividades y consolidado su presencia en varias comunidades autónomas, convirtiéndose en una empresa de referencia en el sector del agua.

La empresa ofrece todos los servicios propios del ciclo integral del agua, desde su captación, potabilización y tratamiento hasta su distribución y posterior recogida y depuración de las aguas residuales. Todas las etapas del ciclo son rigurosamente controladas para limitar el impacto sobre el medio ambiente y garantizar la excelencia en la calidad.

FACSA es también especialista en otras áreas como las aguas industriales, el control de vertidos o los proyectos de ingeniería, reconciliando en todo momento los aspectos económicos, sociales y ambientales para asegurar el desarrollo armónico de todas sus actividades.

Como organización abierta y flexible, FACSA adquiere el férreo compromiso de poner en funcionamiento todos los recursos necesarios para responder a las nuevas exigencias de los clientes y de la sociedad en general, y de adaptarse a los constantes cambios del entorno. Para ello, se fomenta una relación de confianza y transparencia, donde la profesionalidad y cercanía en el servicio facilitado a cada cliente es parte de la cultura de esta empresa.

Como clave de su éxito organizacional, FACSA apuesta por la continua actualización de conocimientos de sus empleados, un equipo integrado por más de 700 profesionales multidisciplinares, para poder ofrecer un servicio de alta calidad enfocado a aportar respuestas y soluciones eficientes.

Las nuevas tecnologías y la inversión en I+D+i se considera una apuesta segura que permite a FACSA perfeccionar y optimizar sus procesos de negocio, facilitando el incremento de la productividad y perfeccionando la relación y conocimiento de sus clientes.

FACSA aporta una gran solvencia tanto técnica como empresarial, derivada de su conocimiento especializado en todas y cada una de las disciplinas que intervienen en el ciclo integral del agua.

1.5.2 Presencia de FACSA en España

FACSA tiene presencia en 9 comunidades autónomas a lo largo del territorio español

1.5.3 FACSA en datos

[3]

Capítulo 2

Análisis descriptivo

Los datos de consumo de agua están contenidos en hojas de cálculo dada su gran envergadura, donde cada fila corresponde a una vivienda.

Estos datos pueden ser medidos de diferente manera, bien trimestral, bimensual o incluso mediante tele lectura, que permite consultar en tiempo real los datos de consumo.

Estas hojas, además de los consumos propiamente dichos, pueden albergar más características que pueden ser importantes en el estudio de la degradación, como la fecha de instalación (día/mes/año), tipo de abastecimiento (vivienda familiar o negocio), si es suministro directo o depósito, lugar y tipo de instalación, zona, etc. En la realización de este proyecto, se ha tenido en cuenta los consumos periódicos y la fecha de instalación, dejando la inclusión de nuevas características para estudios posteriores.

2.1 Datos de partida

Los datos de partida con los que contamos son consumos medios diarios de viviendas en m^3/día en un periodo trimestral.

Estos consumos están tabulados, donde disponemos:

- Descripción y tipo de contador
- Fecha de instalación (días/mes/año)
- Media del consumo antes del cambio con el número de periodos estudiados (normalmente 8) y su incremento
- Media del consumo después del cambio
- Población

A modo de ejemplo podemos observar:

Código de estudio:	
Texto descriptivo:	*CONTADORES*
Fechas cambio de contador:	*desde el 01/01/2007 hasta el 31/12/2007*
Número de periodos a estudiar antes del cambio:	*8*

Fecha de cambio	Media de M3/día antes del cambio	Media de M3/día después del cambio	Incremento	% Incremento	Periodo 1	Periodo 2	Periodo 3	Periodo 4	Periodo 5
09/01/2007	0,24	0,15	-0,08	-35,29 %	0,14	0,25	0,08	0,12	0,18
07/05/2007	0,23	0,15	-0,08	-36,23 %	0,21	0,31	0,37	0,33	0,08
17/01/2007	3,32	3,84	0,52	15,69 %	4,22	4,09	3,43	3,92	4,11
08/02/2007	0,24	0,16	-0,08	-34,98 %	0,26	0,00	0,09	0,15	0,00
08/03/2007	0,01	0,22	0,22	4278,43 %	0,04	0,10	0,01	0,01	0,02
17/01/2007	0,73	0,51	-0,22	-30,55 %	0,62	0,44	0,61	0,73	0,64
06/02/2007	0,52	0,30	-0,22	-41,95 %	0,30	1,08	0,69	0,12	0,79

Fig. 2.1 ejemplo de tabla de consumo. Se han eliminado ciertos datos por privacidad

Algunas características a destacar en esta tabla:

- Cada fila corresponde a la lectura de un único contador y es siempre el mismo.
- Al tratarse de consumos trimestrales, los periodos difieren entre sí 3 meses aproximadamente
- El periodo 1 se corresponde a la primera lectura del contador, donde tendría una vida de 3 meses.
- Los periodos no tienen por qué corresponderse entre sí, todo dependerá de la fecha de instalación del contador. Por ejemplo, en la primera fila, el contador se instaló el 9/01/2007, por lo que su primer periodo pertenece a abril de 2007 (3 meses después)
- A veces, aunque se instale el contador, puede que no se haya dado de alta el servicio por lo que sus primeros periodos marcarían 0.

2.2 Primeros análisis

Si tomamos los consumos de los contadores instalados en 2009 por periodos, mediante una gráfica de barras podemos hacernos a la idea de la cantidad de datos con los que trabajamos.

Observamos que en un principio disponemos de más de 5000 medidas de consumo (es decir, datos de más de 5000 contadores), esta cantidad se va reduciendo bien por cambio de contador o por falta de medición en ese periodo. En el eje de abscisas vemos de qué periodo se trata, dándose una diferencia de 3 meses (aproximadamente) al tratarse de medidas de consumo trimestral.

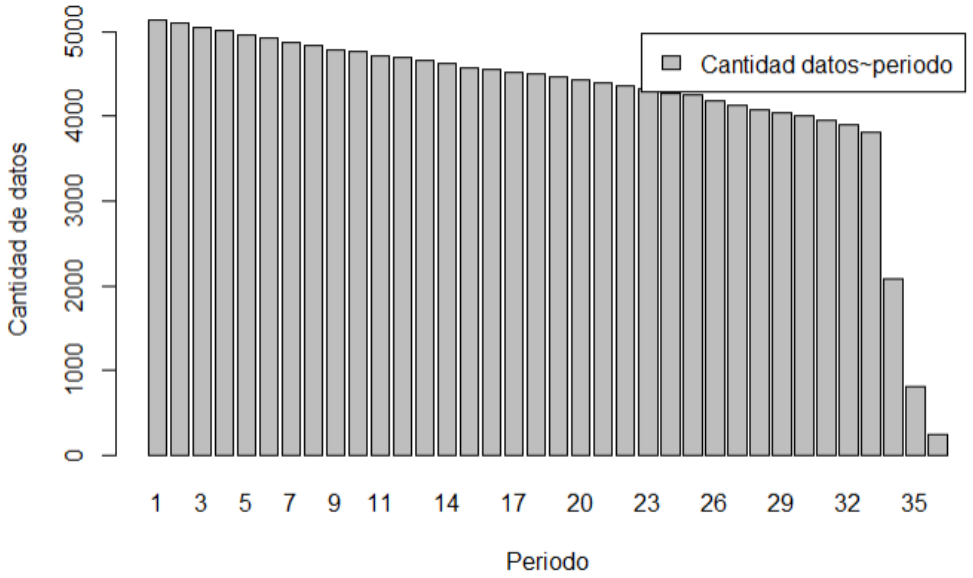


Fig. 2.2 Tamaño de la muestra de consumo del año 2009 por periodo.

Podemos observar también el consumo por periodo mediante diagrama de cajas:

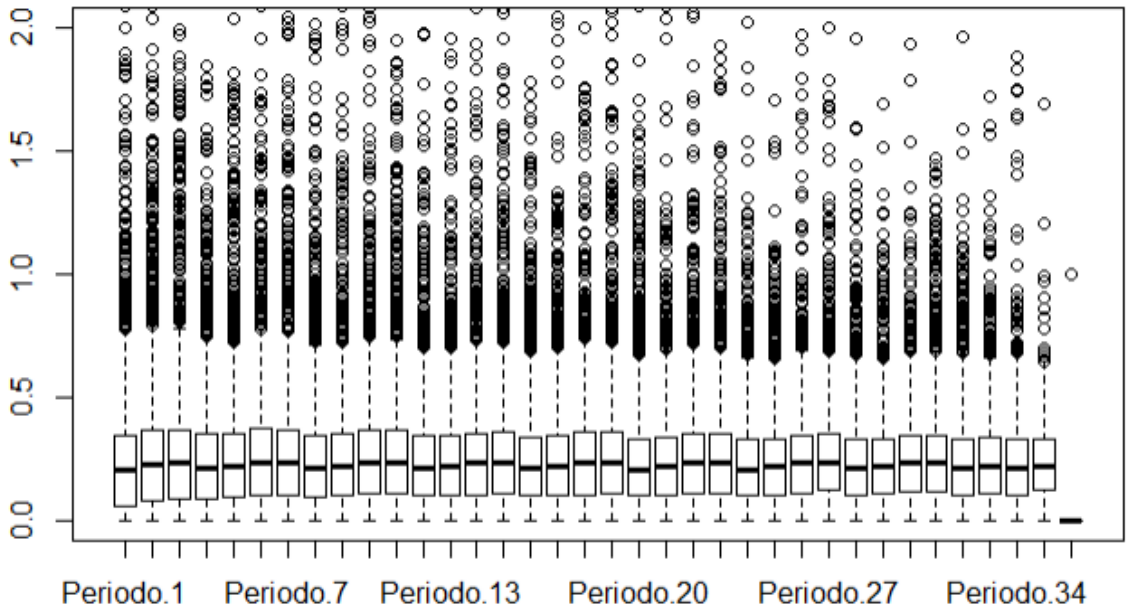


Fig. 2.3 Diagramas de caja para el consumo por periodo de contadores instalados en 2009

Finalmente dibujamos las medias de los mismos periodos:

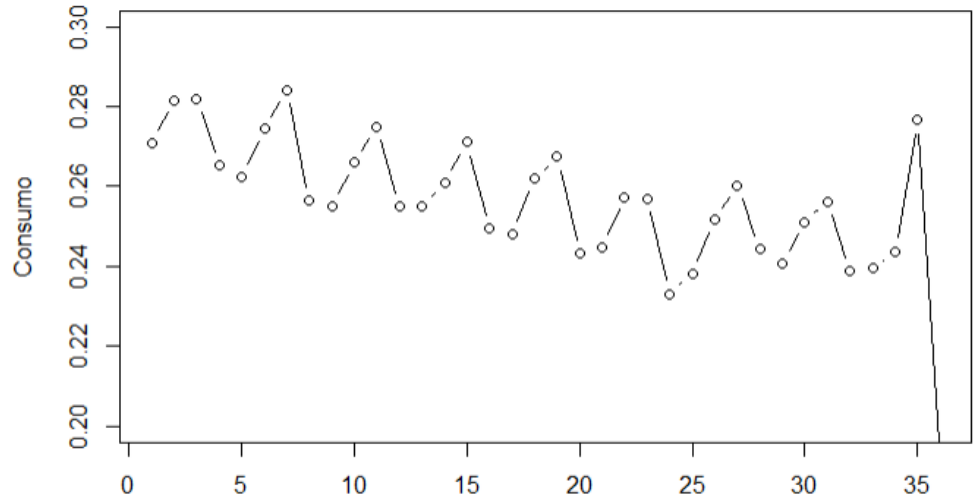


Fig. 2.3 Medias de consumo por periodo de contadores instalados en 2009

Podemos observar también sus correlaciones mediante el coeficiente de Pearson [10] [11]. Para ello, vamos a diferenciar entre los periodos que pertenecen al mismo trimestre de diferentes años:

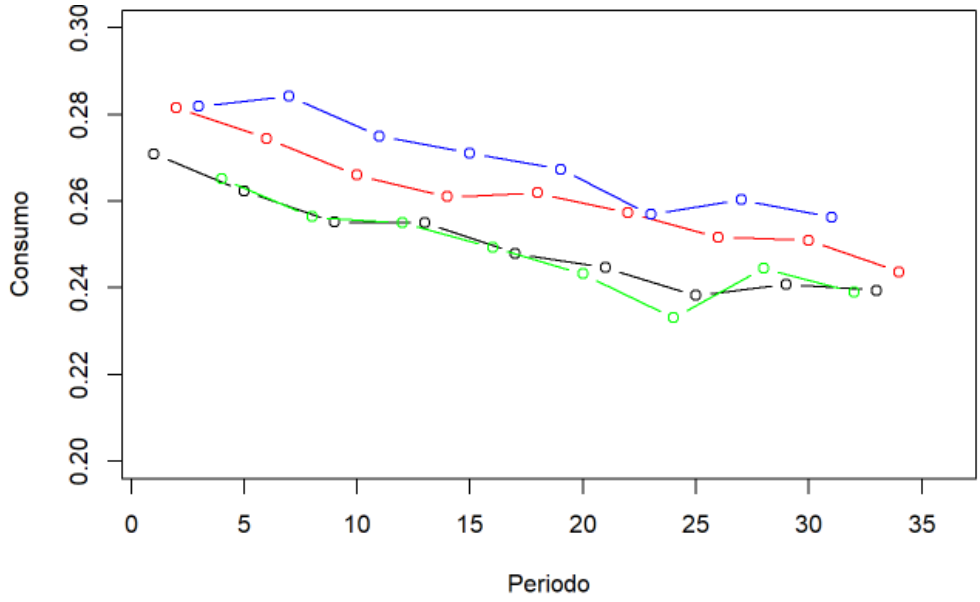


Fig. 2.4 Medias de consumo para mismos trimestres de diferentes de contadores 2009

Observamos la correlación entre los datos, que verificamos con sus coeficientes:

- Primer cuatrimestre: -0.9555669 ***Black***

- Segundo cuatrimestre: -0.9762515 ***Red***

- Tercer cuatrimestre: -0.9595181 ***Blue***

- Cuarto cuatrimestre: -0.8847397 ***Green***

Lo que nos confirma una estacionalidad en los datos de consumo además de una tendencia negativa del mismo.

2.3 Tendencia del consumo

Para ver de manera más cómoda la tendencia de estos consumos podemos definir una serie temporal, trimestral en este caso, y realizamos su estudio según la descomposición aditiva o multiplicativa en 4 factores: [4]

- T_t: tendencia
- E_t: estacionalidad
- C_t: componente cíclica
- I_c: componente aleatoria

Ejemplo)

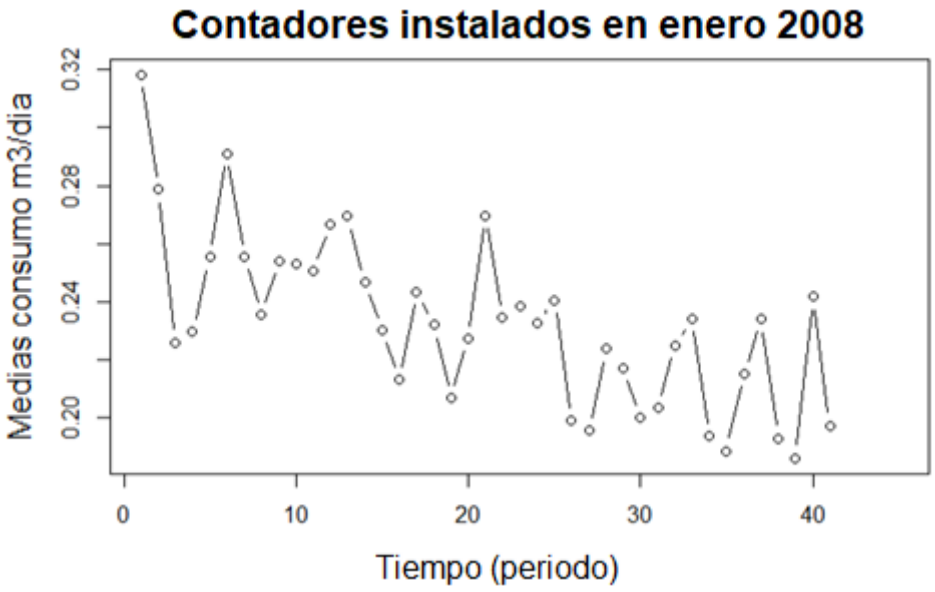


Fig. 2.5 Medias de consumo para diferentes periodos de contadores en 2008

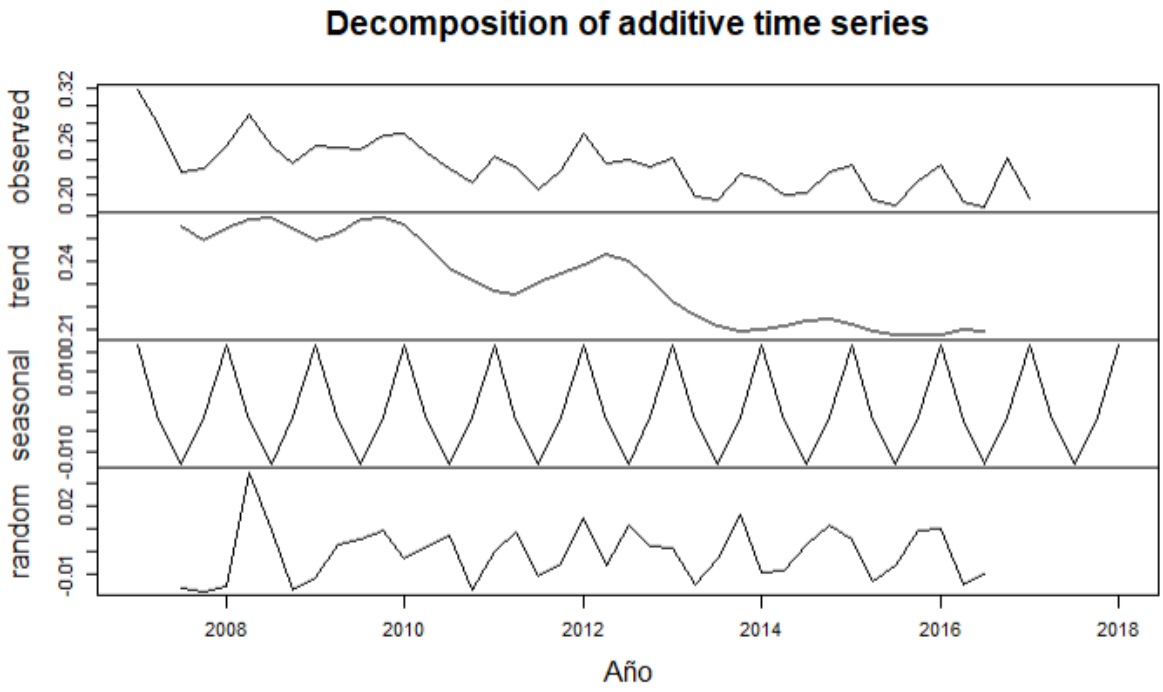


Fig. 2.6 Descomposición de serie temporal contadores 2008

Utilizando en este caso un modelo aditivo por media móvil centrada basada en cuatro periodos, con este método de descomposición tendríamos una tendencia más suave y cercana a los datos, con lo que nos podemos hacer una idea de cómo ha ido variando el consumo con el tiempo de una forma más clara.

Mediante la componente aleatoria, vemos que este modelo aditivo se adapta bien a nuestros datos, ya que esta componente aleatoria es lo que se denomina ruido blanco con media = 0.

Una media móvil es un valor promedio en un número de sesiones determinado, por lo que cada media móvil es la media de un subconjunto de los datos originales.

Podemos diferenciar varias clases de media móvil, como la media móvil simple, donde se hace una media aritmética de los *"n"* datos considerados en el subconjunto, o media móvil ponderada, donde cada dato del subconjunto para realizar la media aritmética está multiplicado por un factor o peso, en el que se le puede dar más importancia a unos datos respecto a otros.

Podemos ver algunos ejemplos:

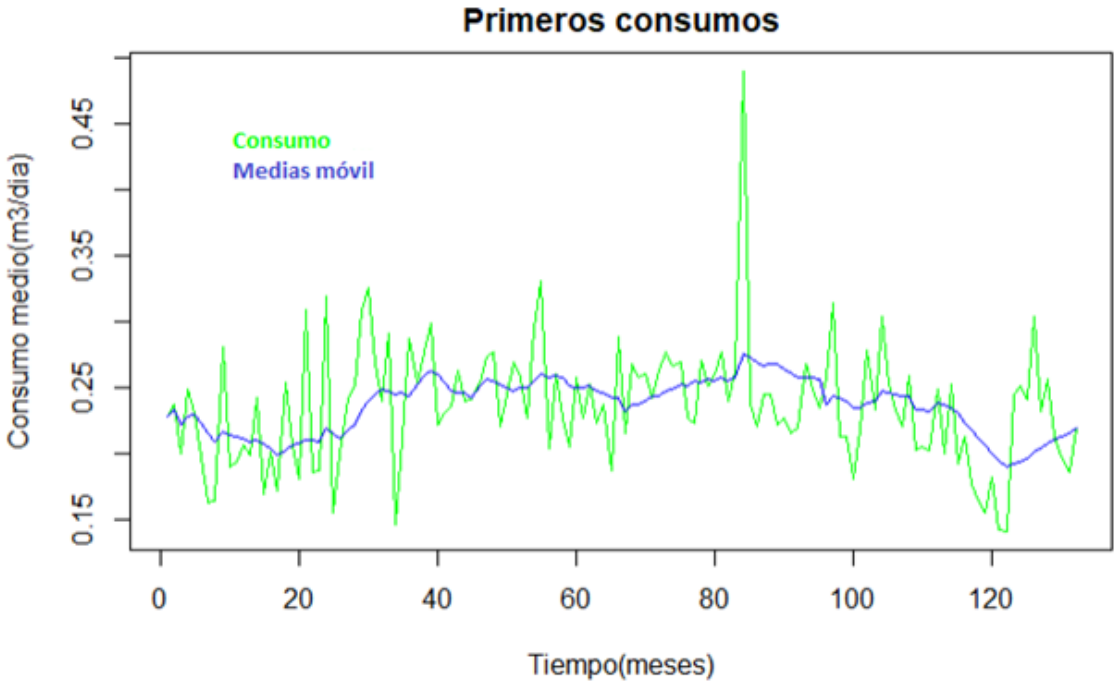


Fig. 2.7 Consumo medio de primeros periodos frente a su media móvil.[1]

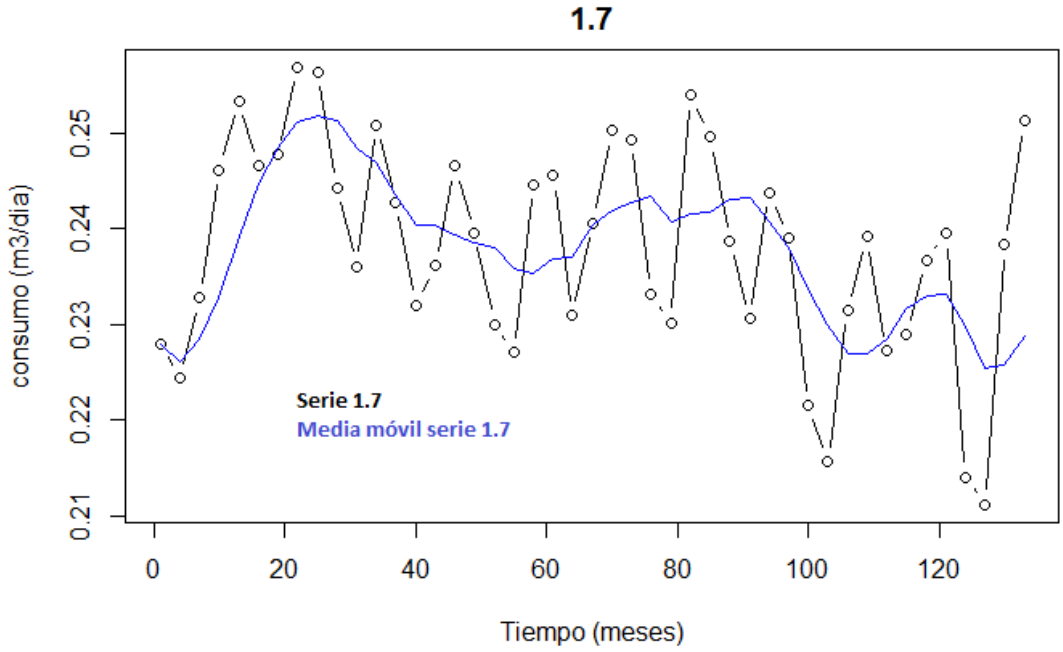


Fig. 2.8 Consumo medio contadores instalados en enero del 2007 y su media móvil.

[1] En esta gráfica se ha tomado media móvil para 12 periodos al tratarse, como veremos más adelante, de medidas mensuales.

2.4 Análisis de normalidad: test de bondad de ajuste

Una parte de los estudios propuestos en este proyecto se basa en la realización de contrastes de hipótesis. Para llevarlos a cabo deberemos saber la distribución de la que provienen los datos.

Más que saber su tipo de distribución, lo que nos interesa realmente es si siguen una distribución normal, en cuyo caso podríamos utilizar estadísticos paramétricos como la t de Student o si, por el contrario, deberíamos emplear estadísticos no paramétricos.

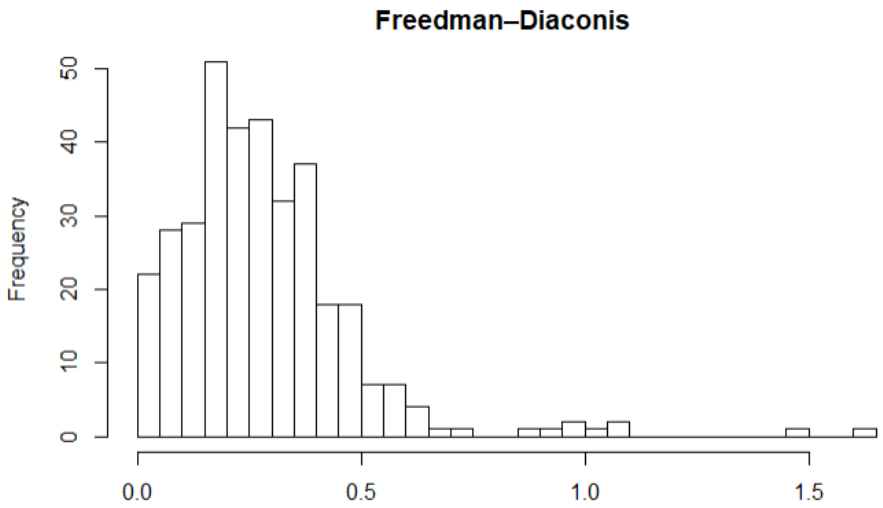
2.4.1 Análisis gráfico

Graficaremos los datos como primera inspección.

Como sus valores están tomados de un conjunto continuo deberemos discretizarlos. Para ello se han seguido las reglas de *Scott[13] (1985), Freedman–Diaconis[15] (FD,1981) y Sturges[14] (1926),* donde el número de clases a considerar para la realización de un histograma sería [12]:

- *Scott*: $c = {3{,}5 * s}/{\sqrt[3]{n}}$ donde c es el número de clases, s la desviación estándar y n el tamaño de la muestra
- *Sturges:* $c = 1 + \log_2(n)$, donde c son las clases y *n* es el tamaño de la muestra.
- *Freedman–Diaconis:* $c = {2 * IQR}/{\sqrt[3]{n}}$ donde IQR = Q3- Q1

Si tomamos ahora el consumo para un periodo y realizamos histogramas según los 3 métodos propuestos obtenemos:



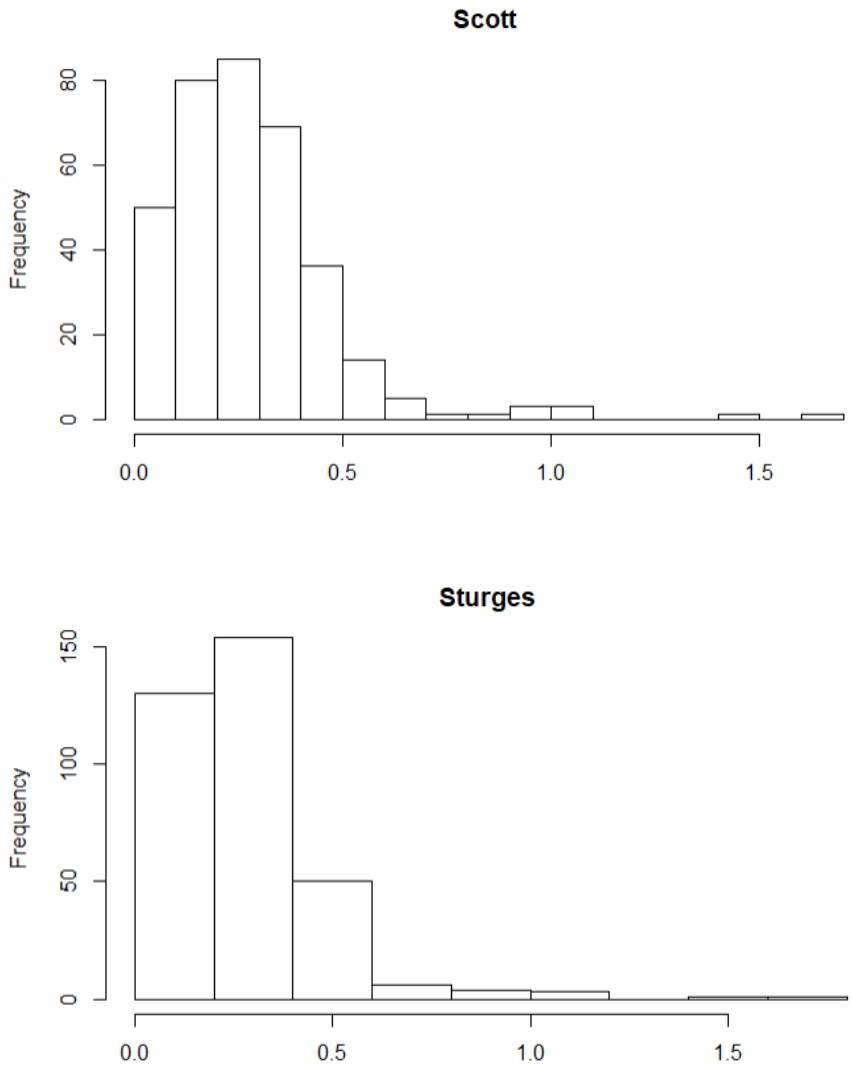


Fig. 2.9 Histogramas sobre datos de consumo utilizando las reglas de discretización según Scott, Freedman–Diaconis y Sturges.

A primera vista no parecen seguir una distribución normal.

2.4.2 Test de bondad de ajuste

Hemos utilizado los siguientes test de bondad de ajuste para comprobar la normalidad de los datos [5]

- *Intervalo intercuantiles*

  Si los datos se distribuyen de manera normal, IQR/s = $Q_3 – Q_1$/sd = 1,3 (aprox)

- *Kolmogorov-Smirnov (con la corrección Lilliefors)*

Cuando esta prueba se aplica para contrastar la hipótesis de normalidad, el estadístico de prueba es la máxima diferencia:

$$D = \text{máx } |F_n(x_i) - F_0(x_i)|$$

Donde

- xi es el i-ésimo valor observado en la muestra (cuyos valores se han ordenado previamente de menor a mayor).
- $F_n(x_i)$ es un estimador de la probabilidad de observar valores menores o iguales que xi
- $F_0(x_i)|$ es la probabilidad de observar valores menores o iguales que xi cuando H0 es cierta.

- *Test de Anderson Darling*

  Utiliza el estadístico A, donde su fórmula determina si los datos $\{Y_1, Y_2 \ldots, Y_N\}$ provienen de una distribución con función acumulativa F

$$A^2 = -N - S$$

  Donde

$$S = \sum_{K=!}^{N} \frac{2K-1}{N} [\ln(F(Y_K) + \ln(1 - F(Y_{N+1-K}))]$$

- *Cramér von Mises*

Se emplea para juzgar la bondad de una función de distribución acumulada F^* comparada con una función de distribución empírica F_N.Se define como

$$w^2 = \int_{-\infty}^{\infty} [F_N(x) - F^*(x)]^2 dF^*(x)$$

Aplicándolo a una única muestra, F^* es la distribución teórica y F_N la empírica

- *Shapiro-Wilks test for normality*

Se usa para contrastar la normalidad de un conjunto de datos. Plantea como hipótesis nula que la muestra $\{X_1, X_2 \ldots X_n\}$ proviene de una población normalmente distribuida.

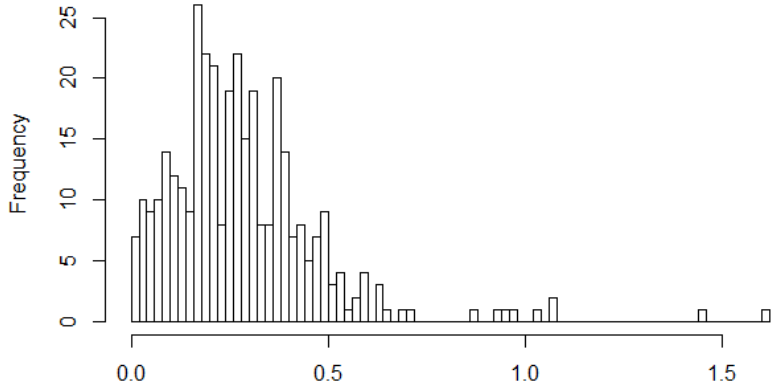
El estadístico del test es

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

2.4.3 Resultados test de bondad de ajuste

- Primera muestra

Nº de datos: 349



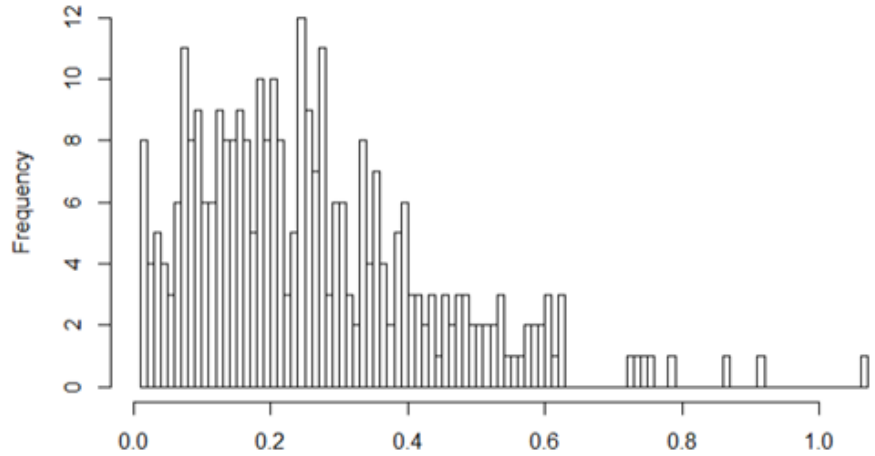

```
(quantile(aux)[4]-quantile(aux)[2])/sd(aux)
lillie.test(aux)$p.value
ad.test(aux)$p.value
cvm.test(aux)$p.value
sf.test(aux)$p.value
```

```
     75%
1.046828
[1] 4.107986e-09
[1] 2.022205e-20
p-value is smaller than 7.37e-10
[1] 2.336102e-16
```

Se rechaza normalidad

- *Segunda muestra*

  Nº de datos: 314




```
(quantile(aux)[4]-quantile(aux)[2])/sd(aux)
lillie.test(aux)$p.value
ad.test(aux)$p.value
cvm.test(aux)$p.value
sf.test(aux)$p.value
```

```
     75%
1.267085
[1] 1.452556e-07
[1] 1.22413e-12
[1] 1.088146e-08
[1] 5.018558e-10
```

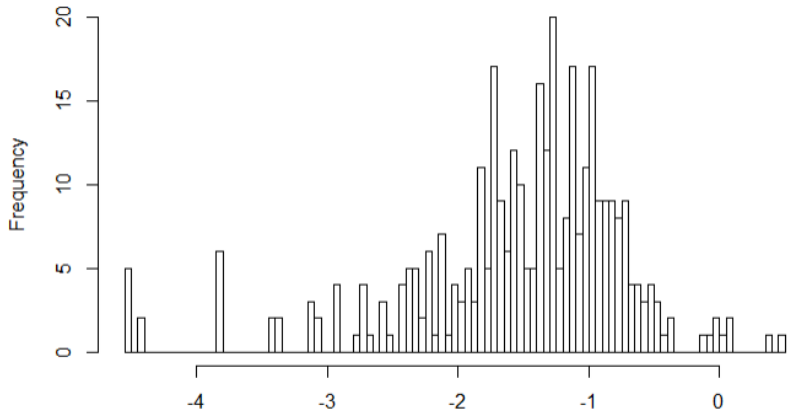
Se rechaza normalidad

2.4.4 Test de bondad de ajuste sobre logaritmos

A veces, cuando los datos están muy cercanos a cero, se pueden tomar los logaritmos de los mismos y puede verse definida una distribución normal, que es lo que comprobamos a continuación.

- Primera muestra

Nº de datos: 349



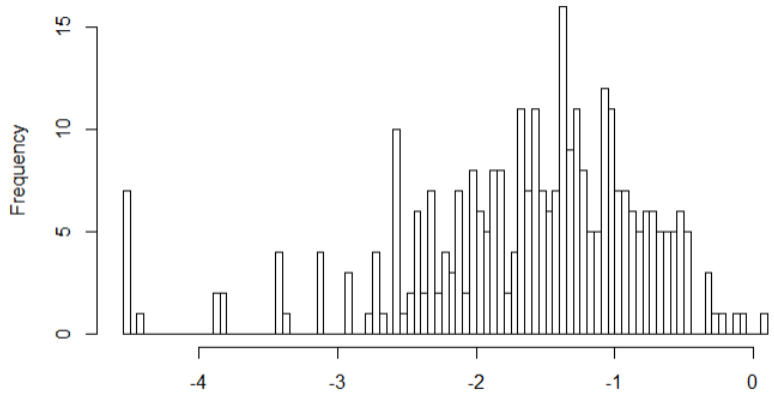

```
(quantile(aux)[4]-quantile(aux)[2])/sd(aux)
lillie.test(aux)$p.value
ad.test(aux)$p.value
cvm.test(aux)$p.value
sf.test(aux)$p.value
```

```
      75%
0.9913613
[1] 8.677822e-16
[1] 7.09497e-21
p-value is smaller than 7.37e-10,
[1] 4.805764e-12
```

Se rechaza normalidad

- *Segunda muestra*

  Nº de datos: 314




```
(quantile(aux)[4]-quantile(aux)[2])/sd(aux)
lillie.test(aux)$p.value
ad.test(aux)$p.value
cvm.test(aux)$p.value
sf.test(aux)$p.value
```

```
      75%
1.162307
[1] 5.939836e-08
[1] 1.041284e-12
[1] 2.252997e-08
[1] 3.705917e-10
```

Se rechaza normalidad

2.5 Análisis por método K-NN: K-Nearest-Neighbor

2.5.1 Introducción

El método K-NN o K vecinos más cercanos es un método de clasificación supervisada donde se cuenta con un conjunto de entrenamiento en el que se conoce la clasificación de cada individuo y un conjunto de estudio donde sus individuos están por clasificarse.

La idea principal de este método es clasificar un dato en la clase más frecuente a la que pertenezcan sus K- vecinos más próximos.

2.5.2 Método K-NN

En la siguiente tabla podemos observar la idea básica de este método

		X_1	...	X_j	...	X_n	C
$(\mathbf{x}_1, c_1)$	1	x_{11}	...	x_{1j}	...	x_{1n}	c_1
	⋮	⋮		⋮		⋮	⋮
$(\mathbf{x}_i, c_i)$	i	x_{i1}	...	x_{ij}	...	x_{in}	c_i
	⋮	⋮		⋮		⋮	⋮
$(\mathbf{x}_N, c_N)$	N	x_{N1}	...	x_{Nj}	...	x_{Nn}	c_N
$\mathbf{x}$	$N+1$	$x_{N+1,1}$	...	$x_{N+1,j}$	...	$X_{N+1,n}$	?

Disponemos de un conjunto $\{(x_1,c_1),\ldots,(x_i,c_i)\ldots,x_N,c_N)\}$ de entrenamiento donde

$$\boldsymbol{x_i} = (x_{1,i} \ldots . x_{i,n}) \; para \; todo \; i = 1, \ldots . N$$

$$\boldsymbol{c_i} \in \{c^1, \ldots, c^m\} \; para \; todo \; i = 1, \ldots . N$$

$c^1 \ldots c^m$ denotan los m posibles valores de la variable clase C. [8]

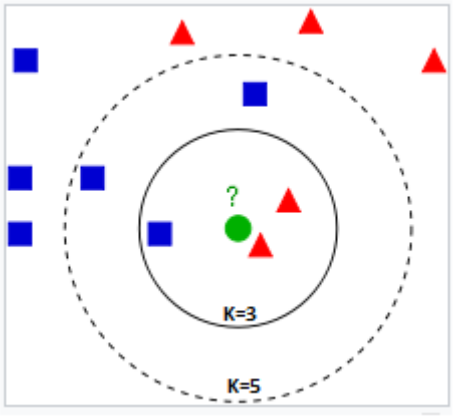
Cada X_i representa una característica de cada individuo y C_i la clase a la que pertenece.

Conocidas las características y la clase de una serie de individuos que forman el conjunto de entrenamiento, queremos predecir o clasificar la clase a la que pertenecen otros individuos que forman el conjunto de estudio.

La distancia entre individuos generalmente se marca con la distancia Euclídea; así, la distancia entre dos individuos (x_1, c_1) y (x_2, c_2) vendrá dada por:

$$d\big((\mathrm{x1},\mathrm{c1}),(\mathrm{x2},\mathrm{c2})\big) = \sqrt{\sum_{j=1}^{n}(x_{1,j} - x_{2,j})^2}$$

Dependiendo del k elegido, se clasificará cada individuo según la clase que predomine entre los k vecinos más cercanos.



En la figura podemos ver la utilización del método K-NN sobre el individuo verde para diferentes k. [9]

Para k=3 clasificaríamos el individuo en la clase roja mientras que para k=5 en la clase azul.

Vemos que es importante entonces determinar la k para poder realizar un mejor estudio.

La elección depende fundamentalmente de los datos. Generalmente, a valores grandes de k se reduce el efecto del ruido en la clasificación, pero crea límites entre clases parecidas, por lo que normalmente se realiza una optimización en uso.

2.5.3 Estudio para 4 muestras en años 2007,2008,2009 y 2010

Haremos un estudio dentro de estos 4 años, para un total de 20 periodos [2]. Hemos tomado las muestras desde el 2007 al 2010 para tener más periodos que incluir en la aplicación de este estudio.

Del año 2007 empezamos por el periodo 13.

Del año 2008 por el periodo 9.

[2] comprendidos entre marzo de 2010 y enero de 2015, 5 años.

Del año 2009 por el periodo 5.

Del año 2010 por el periodo 1.

Eliminaremos aquellas filas que no estén completas, es decir, utilizaremos aquellas viviendas donde tengamos el consumo de todos sus periodos dentro de estos 6 años.

Finalmente tendremos:

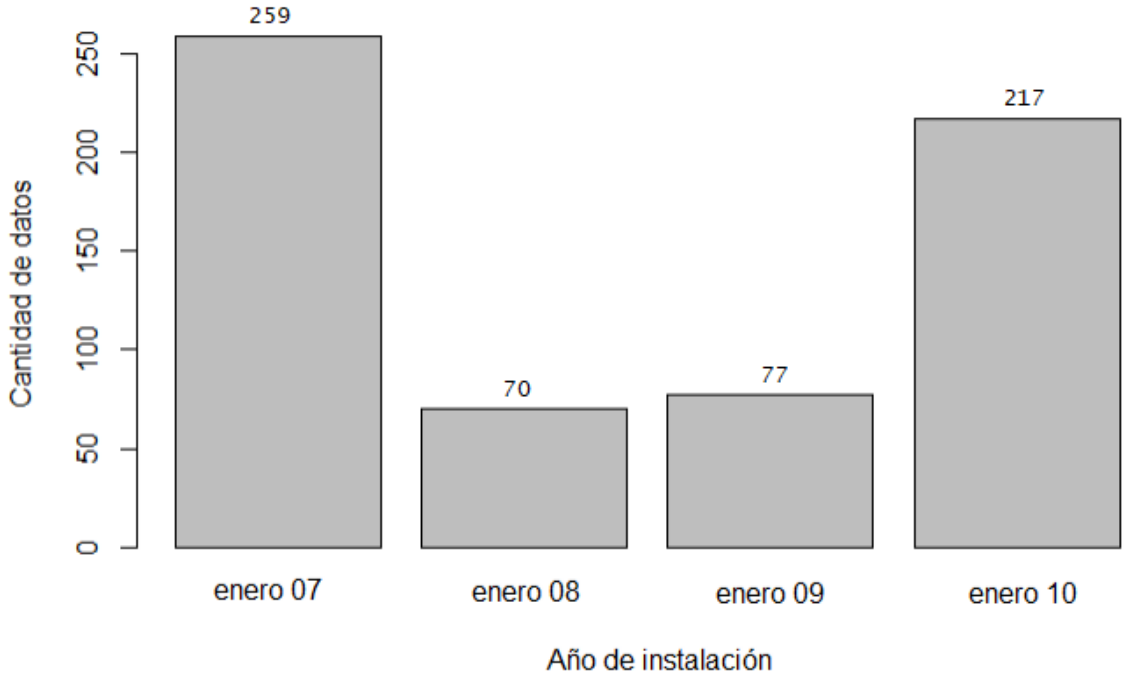


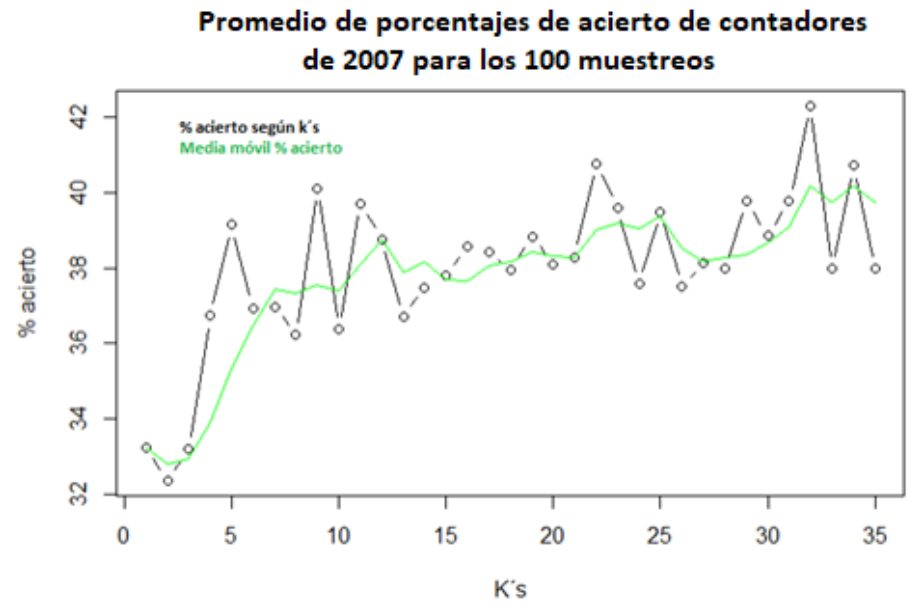
Fig. 2.10 Diagrama de barras de la cantidad de datos en cada muestra según el año de instalación.

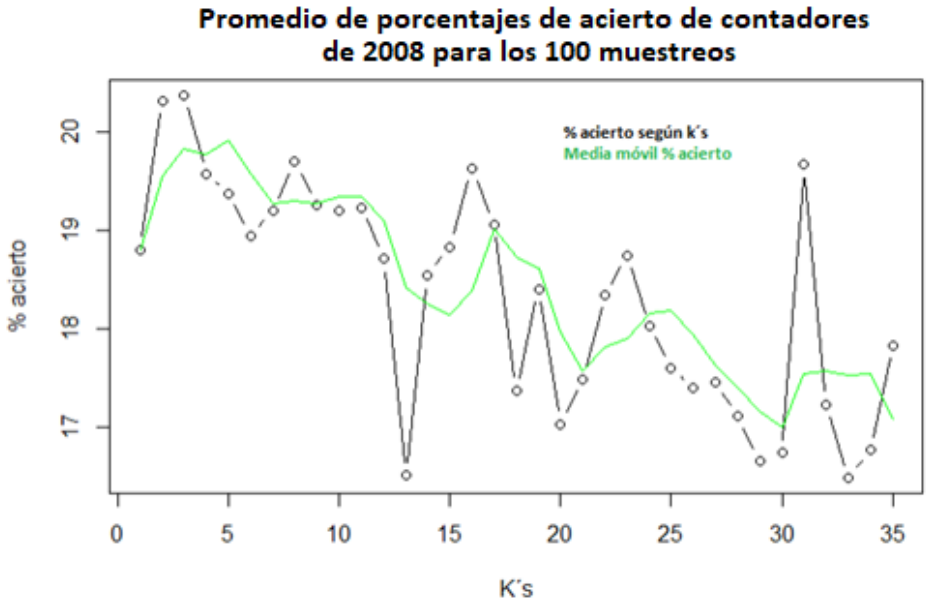
El mínimo son 70 viviendas en enero de 2008; como necesitamos realizar un conjunto de entrenamiento y otro de estudio, tomaremos 35 viviendas de cada año para ambos conjuntos.

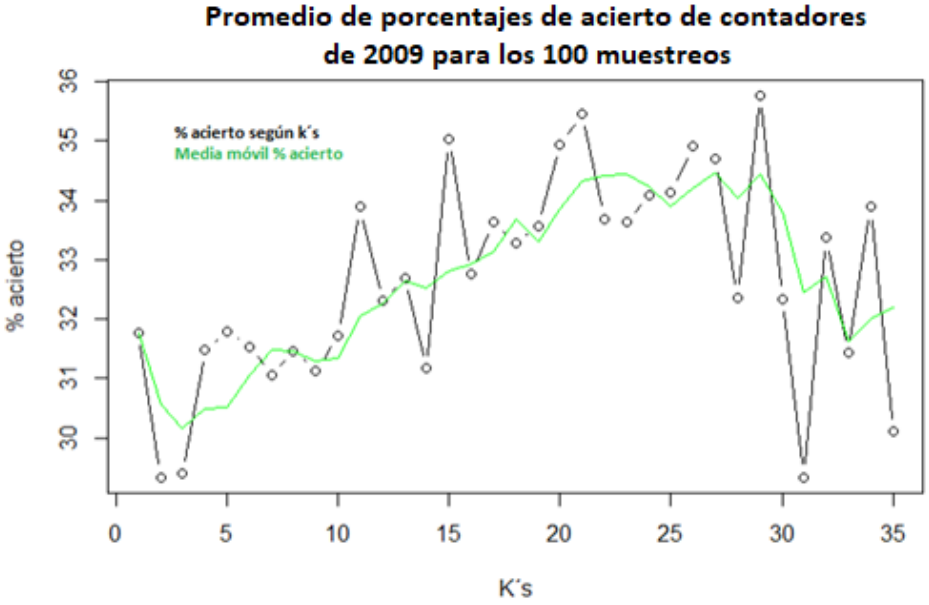
Esta selección la realizaremos aleatoriamente. Vamos a elaborar el estudio para diferentes muestras y diferentes k´s.

Para cada k desde 1 a 35 vamos a seleccionar 100 muestras de tamaño 35 aleatoriamente. Para estas 100 muestras calcularemos la media de aciertos por año y global.

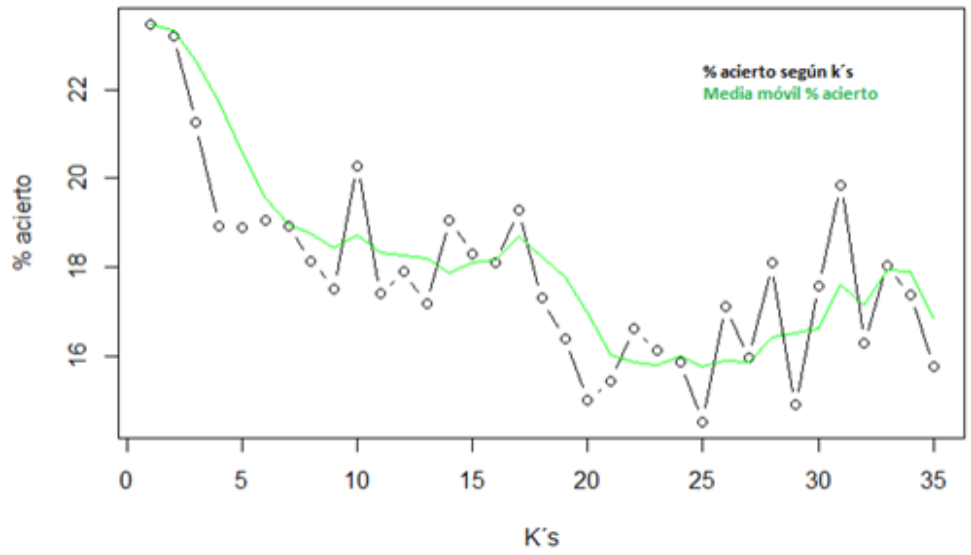
Los resultados se muestran de forma gráfica para los porcentajes de aciertos de cada año según el k seleccionado y una última gráfica con el porcentaje de aciertos total.


Promedio de porcentajes de acierto de contadores de 2007 para los 100 muestreos
% acierto según k´s
Media móvil % acierto
% acierto
K´s

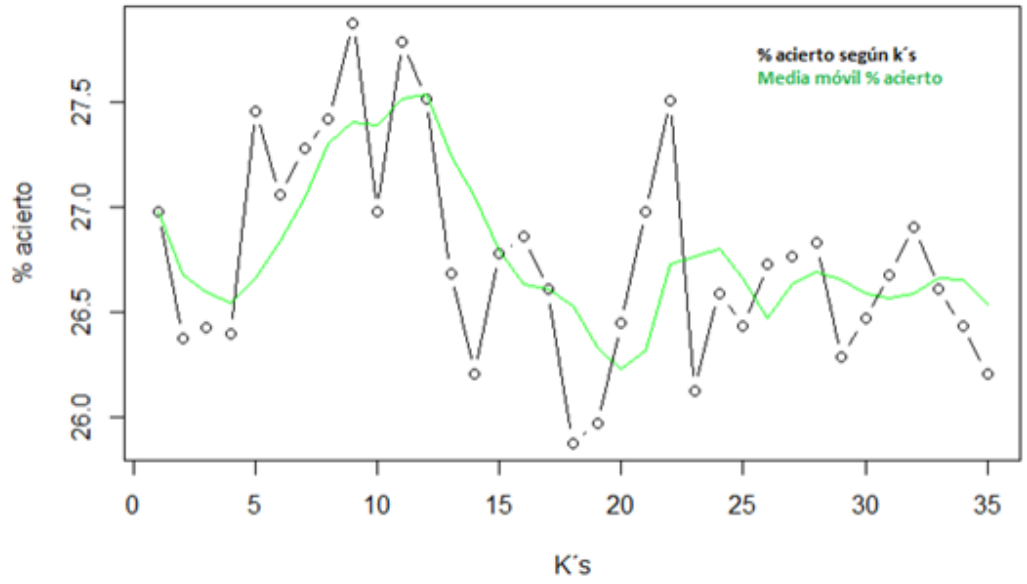

Promedio de porcentajes de acierto de contadores de 2008 para los 100 muestreos
% acierto según k´s
Media móvil % acierto
% acierto
K´s


Promedio de porcentajes de acierto de contadores de 2009 para los 100 muestreos
% acierto según k´s
Media móvil % acierto
% acierto
K´s

Promedio de porcentajes de acierto de contadores de 2010 para los 100 muestreos



Promedio de porcentajes de acierto de contadores de 2007, 2008, 2009 y 2010 para los 100 muestreos



2.5.4 Conclusiones

Observando los resultados, vemos que para los diferentes k´s seleccionados, la muestra del año 2007 aumenta a medida que lo hace k. Además, es la muestra con mayor porcentaje de acierto, comenzando en k=1 en torno al 33 % y terminando el k=35 casi al 40 %.

En la muestra del año 2008 observamos que desciende el porcentaje de acierto a medida que aumentamos k. El porcentaje de acierto está comprendido entre un 20 y un 17%.

Para el año 2009 se tiene un porcentaje relativamente alto, entre un 30 y un 34 % de aciertos. Los porcentajes de aciertos mantienen una línea ascendente hasta llegar a k=25 aproximadamente, que sufre una disminución en el mismo.

Finalmente, en la muestra del año 2010 observamos un comportamiento inverso al año 2009, disminuyendo el porcentaje de acierto en un principio y aumentando de nuevo en torno a k=25. En este caso, a diferencia del año 2009, los porcentajes son menores.

Con lo expuesto hasta ahora vemos que hay un mayor porcentaje de acierto de los contadores instalados en 2007 y 2009 que en el resto y en ambas aumenta a medida que lo hace k. Mientras, en los años 2008 y 2010, tienen menor porcentaje de acierto y van disminuyendo a medida que aumenta k.

Los porcentajes de acierto no parecen ser lo suficientemente altos como para hacer distinciones entre consumos a diferentes edades. Si bien es cierto que para el año 2007 hay un mayor porcentaje de éxito, tampoco es muy alto y, si el porcentaje se diferenciara a mayor edad, para el año 2008 habríamos tenido más éxito que en el 2009 y observamos que no es así.

Así, la diferencia entre las curvas de consumo parece deberse a otras características dentro de las muestras aparte de la edad del contador, como podrían ser contadores de una determinada zona, tipo de abastecimiento o tipo de vivienda suministrada.

Sería interesante poder disponer de más características dentro de cada vivienda de las muestras para poder clasificarlas de una manera más homogénea.

Capítulo 3

Primer método: estudio de la degradación respecto a la edad del contador

3.1 Introducción

Este primer método consiste en un estudio de la degradación del contador según la edad del mismo. Lo que realizamos son sucesivos contrastes de hipótesis entre dos muestras de contadores que difieren en una edad determinada, entendiéndose la misma como los días, meses y años que llevan instalados y en funcionamiento, para verificar si lo marcado en una muestra y otra se puede considerar distinto.

Es por eso que se han clasificado los datos según el año y mes de instalación de su contador. Con ello lo que se intenta es aproximar las lecturas a comparar entre sí a un mismo tiempo donde se verán afectadas por los mismos factores externos, con la salvedad impuesta de su diferencia de edad.

Ejemplo)

Si tomamos como muestra los contadores instalados el 1/07, al ser las medidas de consumo trimestrales, obtendríamos lecturas en:

Mes/año	*1/7*	*2/7*	*3/7*	*4/7*	*5/7*	*6/7*	*7/7*	*…*	*10/7*	*…*	*1/8*	*…*
Consumo	*Inst.*	-	-	*X*	-	-	*X*	*…*	*X*	*…*	*X*	*…*
Periodo				*1*			*2*	*…*	*3*	*…*	*4*	*…*

Por lo que los consumos de los contadores instalados el 1/07 podrán ser comparados temporalmente con los consumos de los contadores instalados donde estén las "X"[3].

Idealmente esta clasificación se realizaría por día/mes/año, pero así se obtienen muy pocos datos en cada muestra, por lo que el rango se puede ampliar por quincenas o por meses, para que, aunque puedan estar un poco más separados entre sí en el tiempo, haya más datos en cada muestra. En nuestro estudio hemos tomado una clasificación mensual.

3.2 Regla empírica para detectar valores anómalos

Hay ocasiones en que un conjunto de datos contiene observaciones inconsistentes. Una observación "*y"* que es inusualmente grande o pequeña en relación con los demás valores de un conjunto de datos, se denomina valor anómalo.

Estos valores por lo general son atribuibles a una de las siguientes causas:

- La determinación se registra incorrectamente.
- La determinación proviene de una población distinta.
- La determinación es correcta, pero representa un suceso poco común o fortuito.

Existen diferentes métodos empíricos para detectar valores anómalos. Los más conocidos son:

3.2.1 Valores de z

El valor de z de un valor *y* de un conjunto de datos es la distancia a la que se encuentra *y* por arriba o por debajo de la media de la muestra $\bar{y}$, medida en unidades de desviación estándar

Con este valor, la regla empírica nos dice que las observaciones con valores de z mayores que 3 en valor absoluto, son valores anómalos. [6]

3.2.2 Intervalo intercuartil: gráficas de cuadro

El intervalo intercuantil, IQR, es la distancia entre los cuartiles superior e inferior

$$IQR = Q_u - Q_L$$

Siendo Q_u cuartil superior y Q_L cuartil inferior, tomaremos como cotas interiores aquellas que se encuentran a una distancia de 1,5 IQR de IQR y cotas exteriores a las que se encuentran a una distancia de 3 IQR

[3] *Las lecturas no son exactamente trimestrales, pueden variar unos días.*

Las observaciones que caen entre las cotas interiores y exteriores se denominan *posibles valores anómalos,* mientras que las observaciones que caen fuera de las cotas exteriores se denominan *valores anómalos muy probables.*

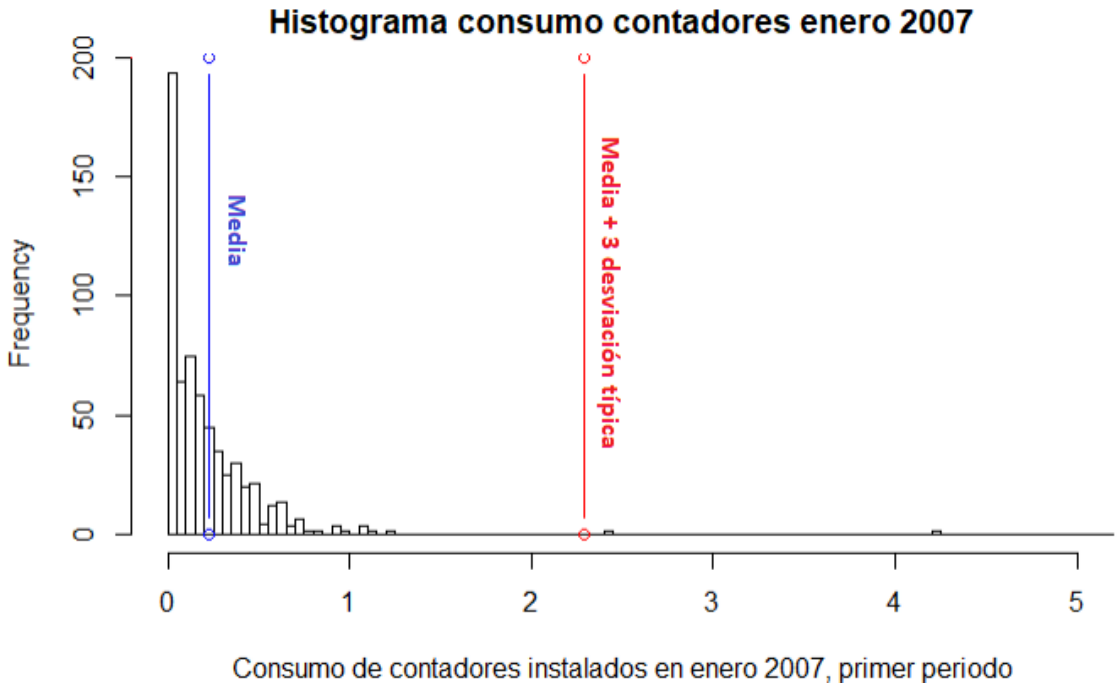
Tanto el método de valor de z como el de gráficas de cuadro establecen límites empíricos fuera de los cuales un valor observado *y* se considera como valor anómalo. Por lo regular, ambos métodos producen resultados similares, sin embargo, la presencia de uno o más valores anómalos en un conjunto de datos puede inflar el valor *s* en que se basa el cálculo del valor de z, por tanto, es menos probable que una observación anómala tenga un valor de z mayor que 3 en valor absoluto. En contraste, los valores de los cuartiles empleados para calcular las cotas de una gráfica no resultan afectadas por la presencia de estos valores. Es por ello que hemos seleccionado el método de valor de z en los sucesivos capítulos.

3.3 Primeros análisis

En primer lugar, vamos a centrarnos en la lectura de los "primeros periodos", es decir, la primera lectura de los contadores recién instalados en sucesivos meses y años (enero 2007, febrero 2007, marzo 2007, …diciembre 2017)

Como los datos se separan por mes de instalación, obtendremos 12 muestras de "primeras lecturas" por año.

En una inspección visual podemos ver datos considerados anómalos. Como ejemplo observamos las muestras de contadores instalados en enero y abril de 2007.



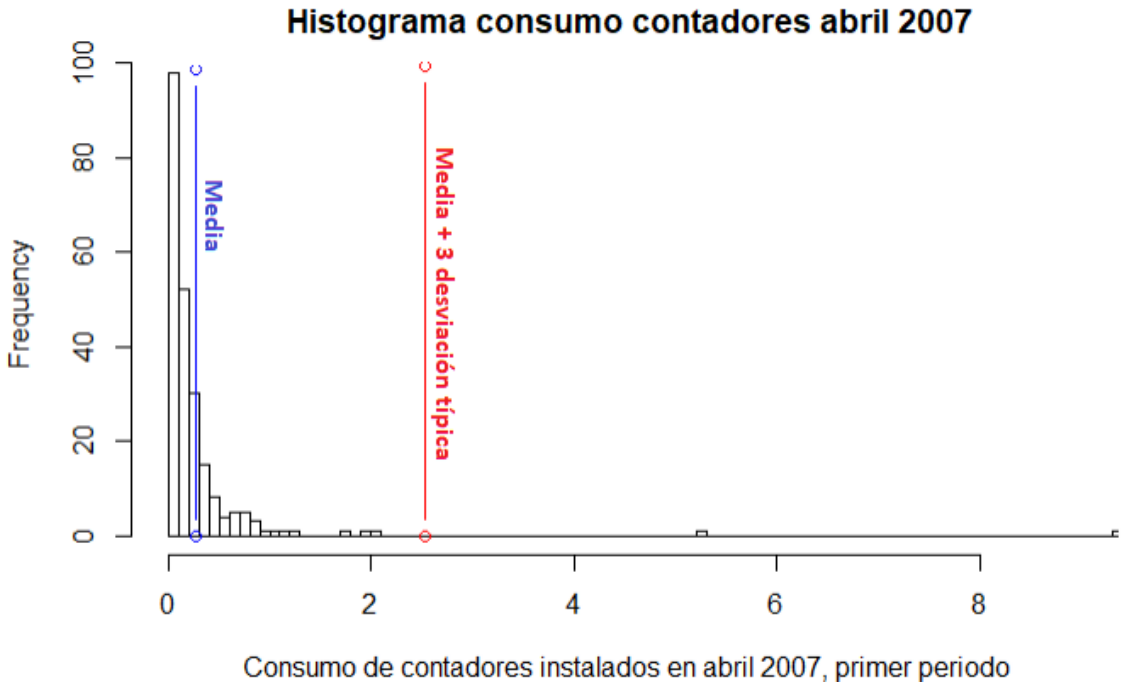
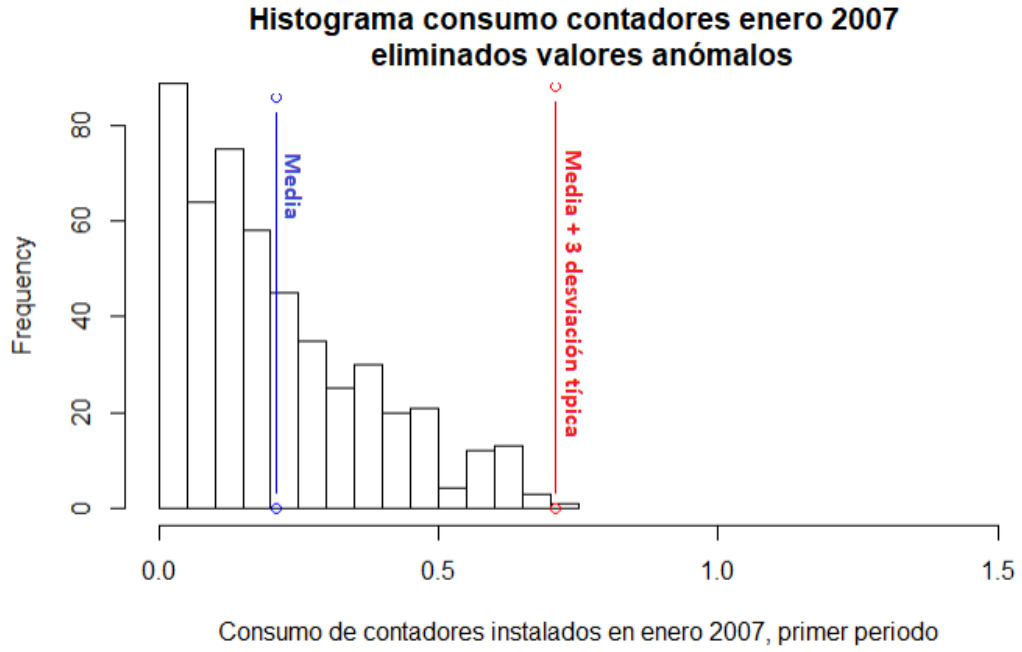


Fig. 3.1 Observamos lo marcado por las muestras de contadores instalados en enero y abril de 2007 en su primer periodo. Marcadas están sus medias y los límites para valores anómalos.

Utilizando sucesivamente la regla empírica por valores de z, eliminamos aquellos valores anómalos. La utilización repetida de esta regla empírica se debe a que pueden existir mediciones "enmascaradas", es decir, que haya en la muestra algunas observaciones que sean anómalas, y que al eliminarlas y calcular de nuevo el valor de z, haya otras que también salgan.

Si eliminamos estos valores fuera de intervalos, podemos observar finalmente como quedarían las anteriores muestras:



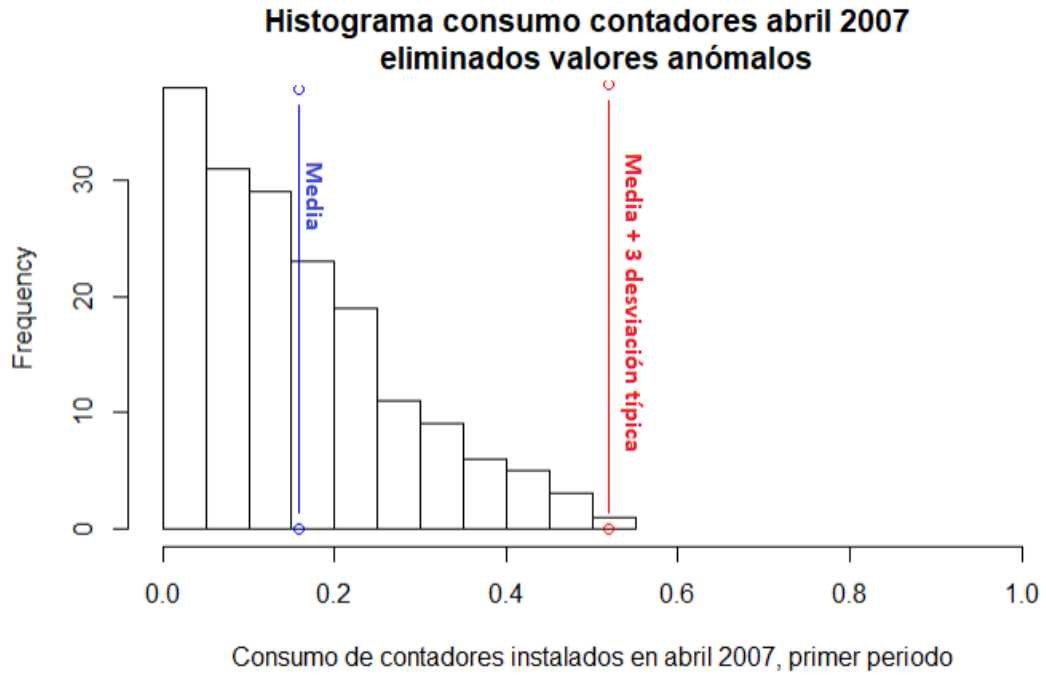


Fig. 3.2 Observamos las muestras de los contadores instalados en enero y abril del 2007 una vez eliminadas las observaciones fuera de intervalo

Si realizamos este proceso con las sucesivas muestras de contadores instalados en diferentes meses, comenzando en enero del 2007, mes a mes, hasta diciembre de 2017, podemos tabular los resultados. A continuación, se exponen los resultados obtenidos desde enero de 2007 hasta agosto de 2007.

Muestra	Medianas	Cantidad de datos
1.7	0.167	496
2.7	0.154	401
3.7	0.152	366
4.7	0.137	175
5.7	0.145	226
6.7	0.000	17
7.7	0.112	101
8.7	0.131	85

Simbología

2.7 =mes. Año, ie, 2.7 serían los contadores instalados en febrero del 2007

Esta mediana de 0.167 m^3/día pertenecería a la mediana del periodo1 de los 496 contadores instalados el 1.7

Las medianas nulas pertenecen a aquellos meses donde hay menos de 30 datos. Tomamos 30 ya que estadísticamente sería considerada una muestra grande y son requeridas en pasos posteriores.

Según el estudio que hemos realizado, los datos no siguen una distribución normal, por lo que es preferible utilizar la mediana a la media, ya que es una medida central más robusta y nos informará mejor sobre la muestra al tener en ella valores muy dispares, lo cual afectaría en mayor medida a la media que a la mediana.

Si graficamos esta mediana de todos los primeros periodos obtenemos:

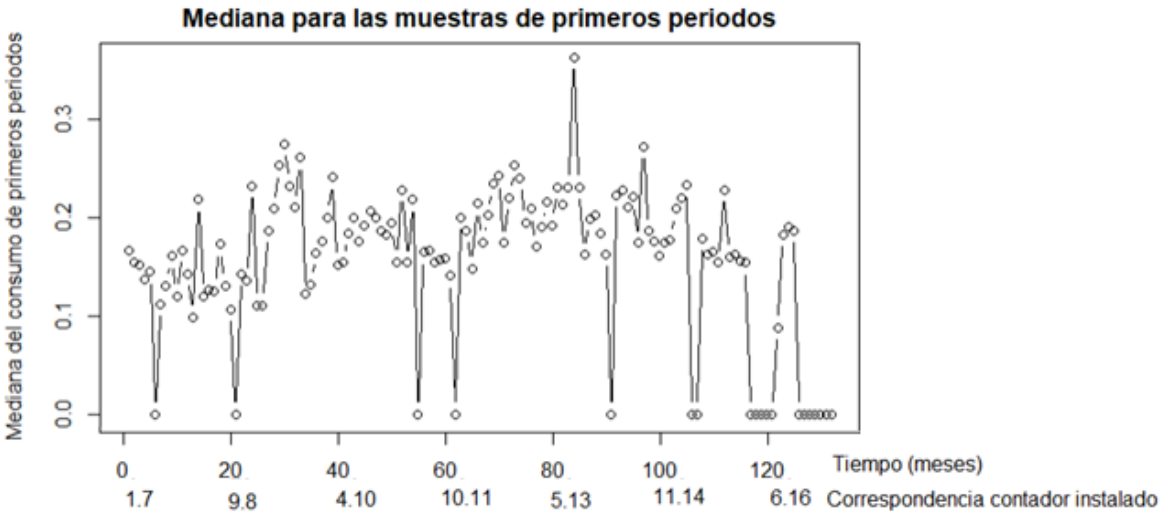


Fig. 3.3 Observamos las medianas del consumo del primer periodo de contadores instalados en diferentes meses, empezando por enero del 2007 hasta diciembre de 2017. También vemos la correspondencia entre meses y periodos.

El tiempo en meses expuesto en el eje de abscisas va ligado al mes de instalación. Es decir, empezando desde enero del 2007, en el mes 1, se instalaron los 1.7, en el mes 2 se instalaron los 2.7… así sucesivamente.

Los valores puestos a cero representaban muestras de menos de 30 datos. Simplemente de manera gráfica, podemos promediar los consumos que no tenemos mediante:

- Si el valor nulo es de una muestra dentro del año 2007 (como ocurre con los 6.7), tomamos la mediana entre el valor anterior y el posterior a ese valor (mediana (5.7) + mediana (7.7) / 2)
- Si el valor nulo está en un año diferente al 2007, tomamos la medida perteneciente al mismo periodo del año anterior.

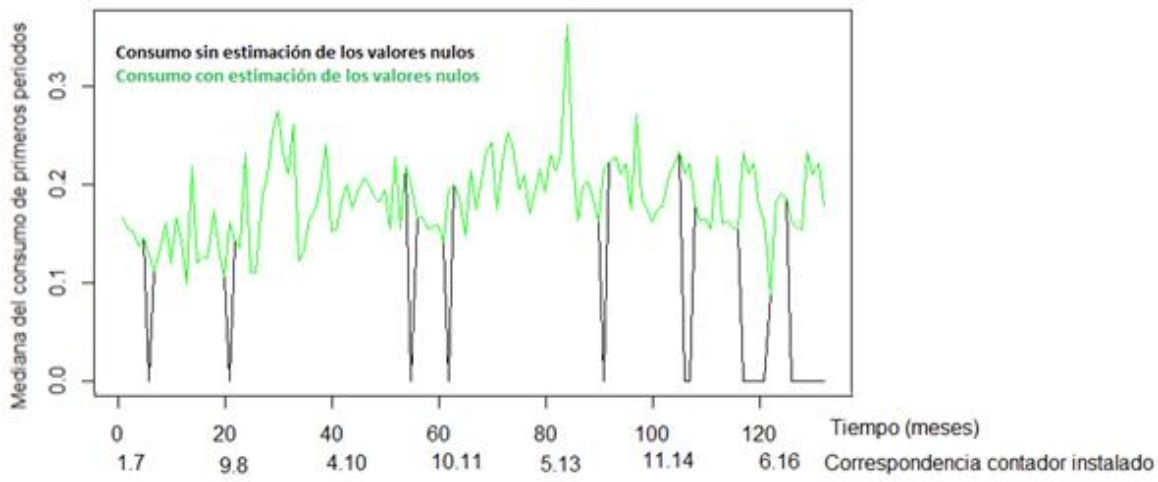


Fig. 3.4 Observamos las medianas de consumo donde hemos superpuesto en verde la misma gráfica de medianas de consumo para los primeros periodos, pero estimando los valores nulos.

También podríamos obtener su tendencia por media móvil

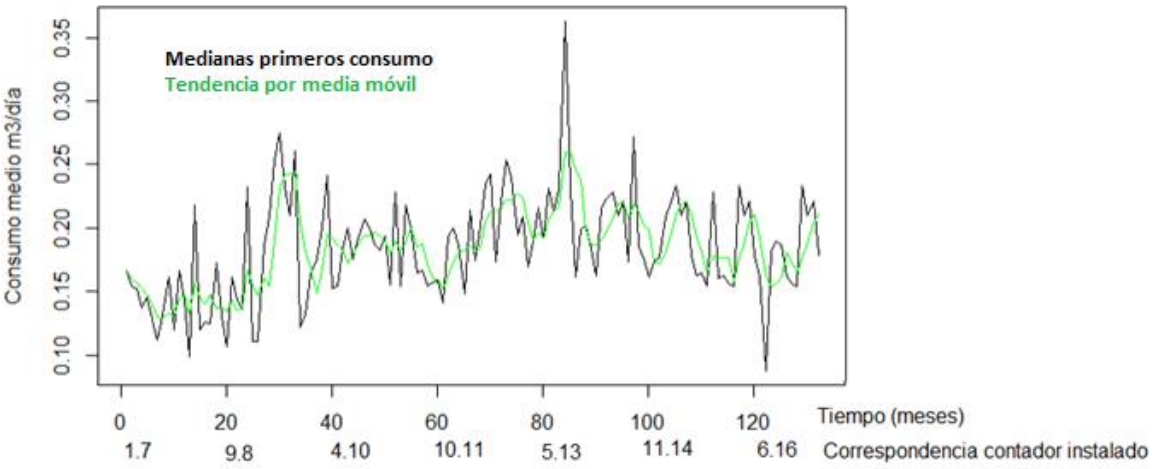


Fig. 3.5 En negro tenemos las medianas de los primeros periodos de muestras de contadores instalados por mes. En verde se realiza la media móvil.

Lo mismo que hemos realizado en el apartado anterior, podemos aplicarlo a las series de consumo clasificadas por mes y año. Es conveniente este tipo de análisis visual, ya que tratamos con una cantidad de datos considerable y este recurso nos facilita la labor.

Como ejemplo, el consumo de los contadores 1.7, los instalados en enero del 2007.

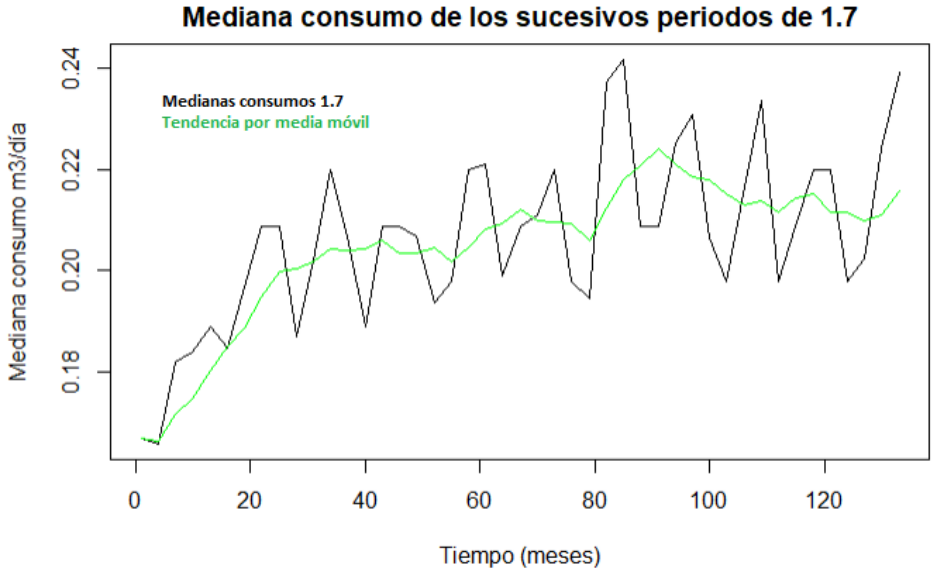


Fig. 3.6 En negro tenemos las medianas de los diferentes periodos para los contadores instalados en enero de 2007. En verde su media móvil.

Si, por ejemplo, sobre la gráfica de primeros periodos trazamos las medianas de los sucesivos periodos de 1.7 podemos observar que la mediana del consumo de los contadores instalados en enero del 2007, no parece que esté por debajo de lo marcado por contadores nuevos.

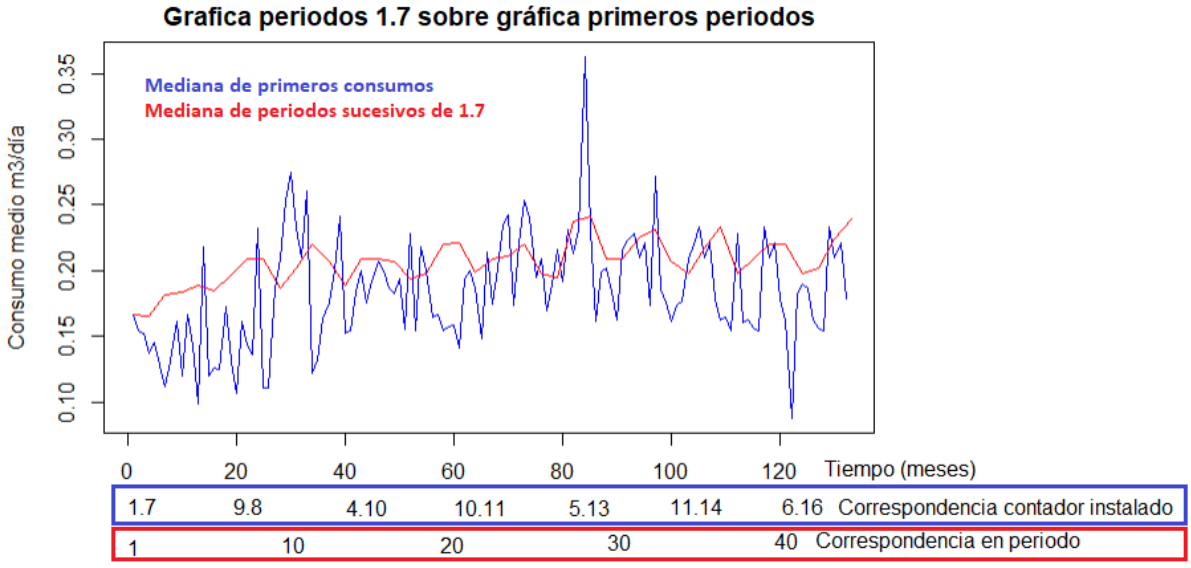


Fig. 3.7 Se grafica en azul la mediana de primeros periodos de muestras de contadores instalados en sucesivos meses. En rojo la mediana de las muestras de sucesivos periodos para contadores instalados en enero del 2007

3.4 Comparación de consumos: U de Mann-Whitney

Con lo realizado anteriormente, podemos hacernos una idea de cómo ha variado gráficamente lo marcado por los contadores instalados en diferentes periodos respecto a lo registrado por los nuevos, pero, ¿existe una diferencia real?

En la siguiente tabla se expone por columna los meses del año y por fila contadores instalados en diferentes meses.

Mes instalación/mes lectura	*1/7*	*…*	*4/7*	*…*	*7/7*	*…*	*10/7*	*…*	*1/8*
1.7	*Instalación*	*…*	*1º lectura*	*…*	***2º lectura***	*…*	***3º lectura***	*…*	***4º lectura***
4.7		*…*	*Instalación*	*…*	***1º lectura***	*…*	*2º lectura*	*…*	*3º lectura*
7.7		*…*		*…*	*instalación*	*…*	***1º lectura***	*…*	*2º lectura*
10.7		*…*		*…*		*…*	*Instalación*	*…*	***1º lectura***

Por ejemplo, vemos que para la fila 1.7 tenemos marcado “instalación” en la columna 1/7 y su primera lectura el 4/7.

Ahora bien, si tomamos la muestra de 2º lectura de los 1.7 y la muestra de 1º lectura de los 4.7, estas muestras coinciden en el tiempo y se verán afectadas por los mismos agentes externos (condiciones atmosféricas, campañas concienciación, etc.) pero los 1.7 tienen ya una edad de 6 meses y los 4.7 una edad de 3 meses. Con ello queremos comparar si lo marcado por estas dos muestras difieren y averiguar cuánto.

Simbología

Llamaremos “nuevo” a aquella muestra de contadores instalados donde tratemos su primer periodo.

Llamamos “antiguo” a aquella muestra de contadores instalados donde tratemos algún periodo distinto al primero.

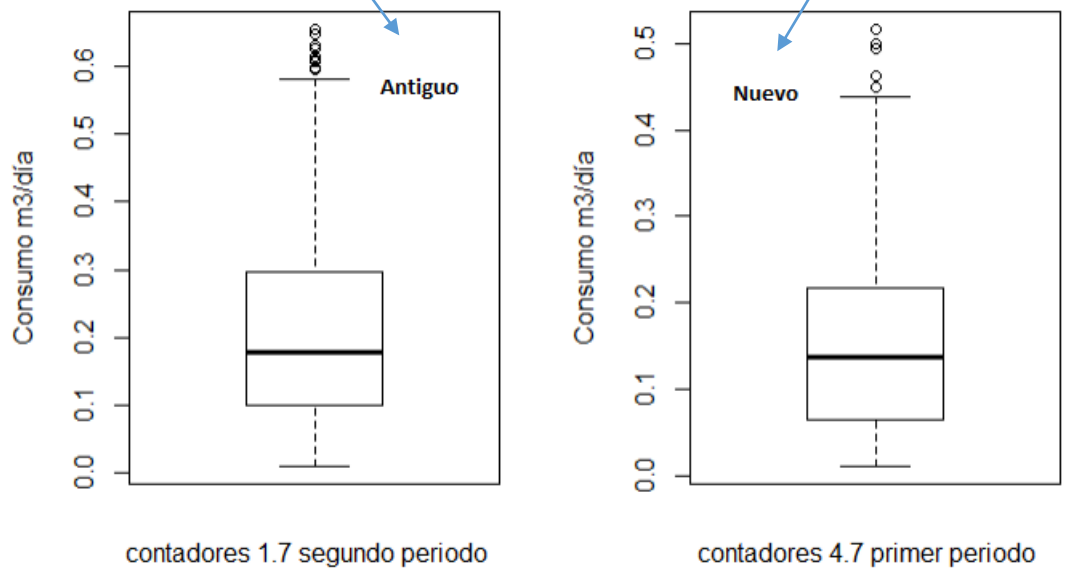


Fig. 3.8 Boxplot para los consumos de 1.7 y 4.7.

¿Podríamos decir que difieren las dos muestras realmente?

Utilizaremos el *contraste de hipótesis* para poder comparar muestras y observar si realmente hay evidencias de que provienen de poblaciones diferentes, en este caso nos interesamos por la mediana.

Para saber qué estadístico utilizar hay que estudiar la procedencia de los datos, es decir, su distribución. Por los diferentes test de bondad de ajuste vimos que no provienen de una normal, por lo que utilizaremos estadísticos no paramétricos, la prueba U de Mann-Whitney sería el equivalente a la t de Student.

Esta prueba es una prueba no paramétrica aplicada a dos muestras independientes. Propuesta inicialmente por Frank Wilcoxon en 1945 para muestras de igual tamaño y extendido a muestras de tamaño arbitrario por Henry B. Mann y D.R. Whitney en 1947.

El planteamiento de partida en esta prueba es: [7]

- Las observaciones de ambos grupos son independientes.
- Las observaciones son variables ordinales o continuas.
- Bajo la hipótesis nula, la distribución de partida de ambos grupos es la misma: P(X > Y) = P(Y > X)
- Bajo la hipótesis alternativa, los valores de una de las muestras tienden a exceder a los de la otra: P(X > Y) + 0.5 P(X = Y) > 0.5.

El estadístico U de contraste se obtiene:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Donde n es el tamaño muestral, R el rango de las muestras. El estadístico U se define como el mínimo entre U_1 y U_2.

La aproximación a la normal cuando tenemos muestras grandes viene dada por la expresión

$$z = (U - m_U)/\sigma_U$$

donde m_U es la media y σ_U la desviación estándar de U dadas por

$$m_U = n_1 n_2/2$$

$$\sigma_U = \sqrt{\frac{n_1 n_2(1 + n_1 + n_2)}{12}}$$

A continuación, con sucesivos contrastes de hipótesis, se compara los consumos de los contadores recién instalados (nuevos) con los consumos registrados por contadores instalados en cierto mes (antiguos), siempre para un mismo tiempo en ambas muestras.

Si en el contraste de hipótesis (con nivel de significación α=0.05) se rechaza la hipótesis nula (que sería de medianas iguales) obtendremos la diferencia de medianas y el porcentaje de esta diferencia en cuanto a incremento o disminución en la mediana según:

$$\% \, dif = \frac{Medinana \; periodos \; 1.7 - mediana \; 1º \; periodos \; (\; 4.7, 7.7..)}{mediana \; 1º \; periodos \; (\; 4.7, 7.7..)} * 100$$

Ejemplo) Comparamos los contadores instalados en enero del 2007 vs contadores instalados posteriormente

Contadores instalados en enero 2007 vs posteriores

1.7 vs	Dato antiguo	Dato nuevo	p-valor	Ma - Mn	% diferen.
4.7	451(periodo 2 de 1.7)	175 (periodo 1 de 4.7)	8.23e-05	0.04	29.5
7.7	430(periodo 3 de 1.7)	101 (periodo 1 de 7.7)	1e-05	0.08	67.13
10.7	415(periodo 4 de 1.7)	152 (periodo 1 de 10.7)	0	0.08	64.49
1.8	394(periodo 5 de 1.7)	111 (periodo 1 de 1.8)	0	0.1	104.4
4.8	385(periodo 6 de 1.7)	213 (periodo 1 de 4.8)	0	0.06	47.9
7.8	386(periodo 7 de 1.7)	92 (periodo 1 de 7.8)	1.3e-06	0.08	57.61
10.8	367(periodo 8 de 1.7)	78 (periodo 1 de 10.8)	2.62e-05	0.07	49.41

En esta tabla tenemos por columnas:

- *Mes/año comparación (4.7 son los contadores instalados en abril de 2007, 7.7 en julio de 2007, 10.8 en octubre de 2008…)*
- *Cantidad de datos de los contadores 1.7 en periodos sucesivos (datos antiguos)*
- *Cantidad de datos de los contadores instalados posteriormente en su primer periodo (datos nuevos)*
- *P-valor por la prueba U de Mann-Whitney*
- *Mediana(datos1.7) – mediana (datos actuales)*
- *Porcentaje de error*

Las comparaciones donde el p-valor sea superior a 0.05 nos informan que no se rechaza la hipótesis nula, por lo que no se puede decir que las muestras difieran en mediana. También puede haber un "fdd" que representa falta de datos, i.e, < 30 datos en alguna de las muestras a comparar.

Graficando la diferencia de medianas y el % de diferencia podemos observarlo de manera más clara:

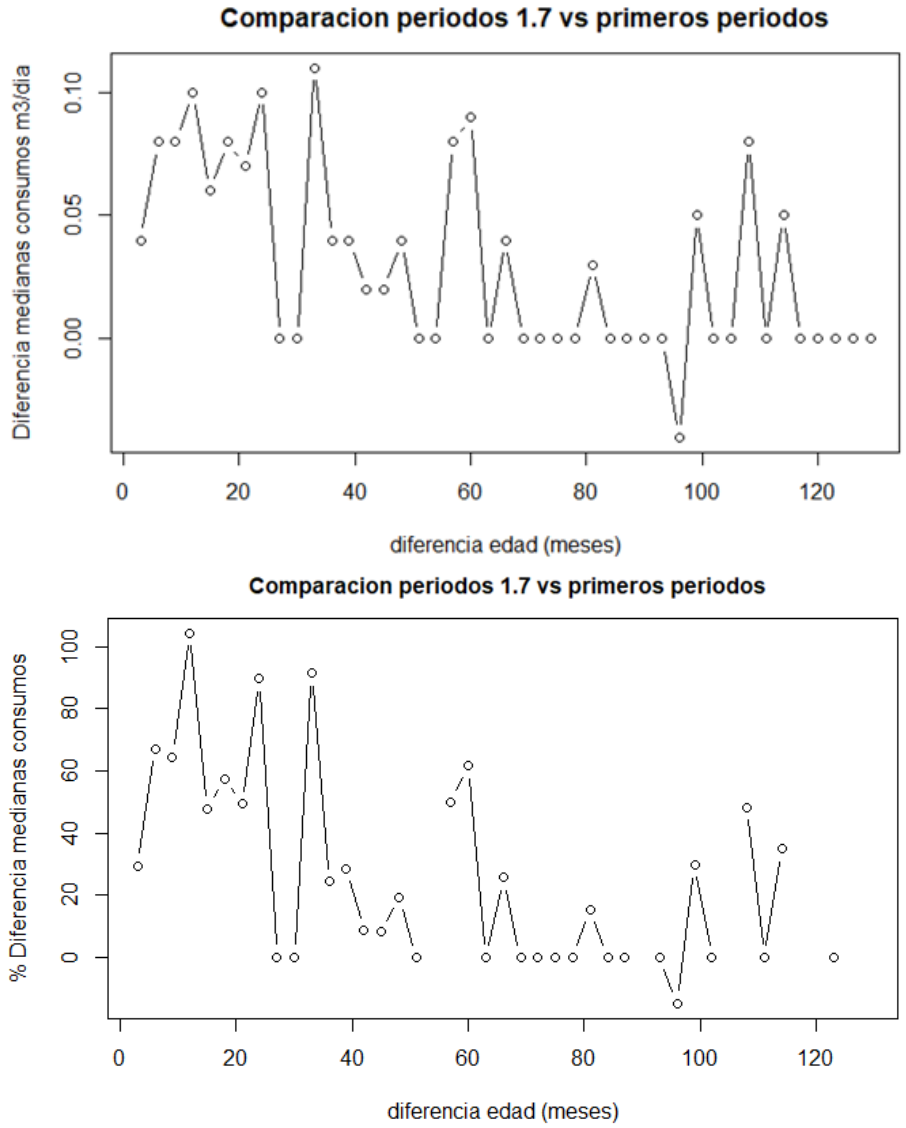


Fig. 3.13 Diferencia de medianas y su porcentaje entre los sucesivos periodos de 1.7 frente al primer periodo de contadores instalados recientemente (nuevos).

3.5 Aplicación sobre datos de consumo trimestral

En este apartado vamos a realizar la aplicación de este primer método de forma completa para los contadores instalados en enero del 2007, es decir, los 1.7.

Como los datos de partida son lecturas trimestrales, para los 1.7 los datos de lectura corresponderán a abril, julio, octubre de 2007, y seguirán con enero, abril, julio y octubre de 2008 y así sucesivamente, esto es, los meses donde nos llegaría la factura a casa.

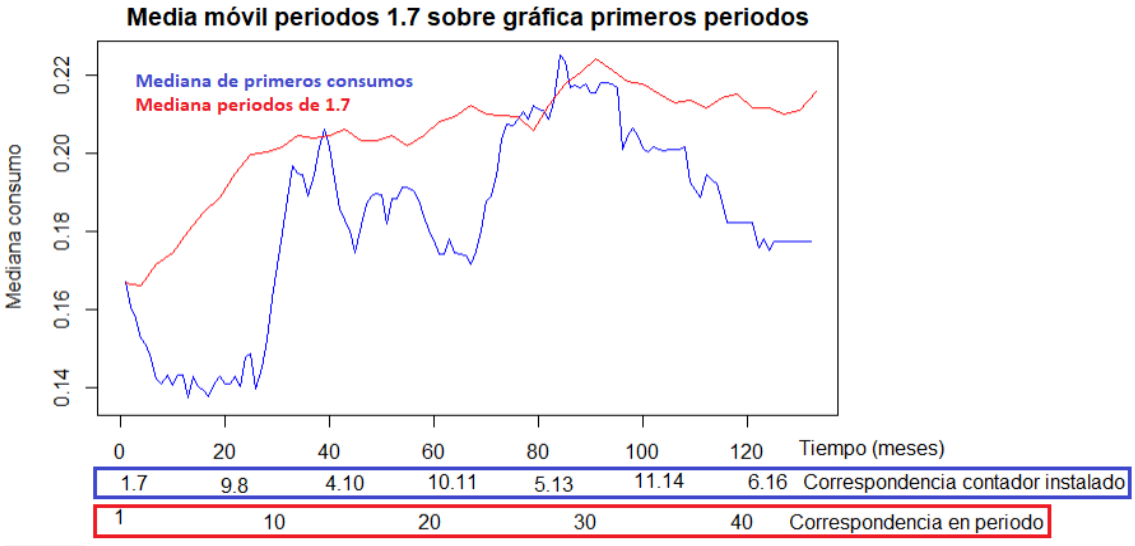
Ya vimos que estas lecturas se corresponden en tiempo a las lecturas de contadores instalados posteriormente en esos mismos meses, como podemos ver en la siguiente tabla:

Mes instalación/mes lectura	*1/7*	...	*4/7*	...	*7/7*	...	*10/7*	...	*1/8*
1.7	*Instalación*	...	*1º lectura*	...	***2º lectura***	...	***3º lectura***	...	***4º lectura***
4.7		...	*Instalación*	...	***1º lectura***	...	*2º lectura*	...	*3º lectura*
7.7		...		...	*instalación*	...	***1º lectura***	...	*2º lectura*
10.7		...		...		...	*Instalación*	...	***1º lectura***

Estas “primeras lecturas” serán comparadas con las lecturas correspondientes de los 1.7, viendo así la diferencia en mediana de las muestras comparadas.

Superponiendo las medianas de “primeras lecturas” con las medianas de las lecturas sucesivas de 1.7, obtenemos una gráfica para un primer análisis visual como vimos en la *figura 3.7*

Este primer análisis visual se intenta mejorar aislando la tendencia por media móvil.



Al utilizar media móvil, en este caso cada 4 periodos para los sucesivos periodos de 1.7, y 12 periodos para las sucesivas muestras de “primeras lecturas”, tendríamos una falta de datos al principio y al final. Para compensarla, lo que se ha realizado es una media entre los datos extremos y situarlos en cada posición, y utilizar el valor del primer dato. No sería necesario, pero mejora el análisis visual.

Realizando las sucesivas comparaciones y tabulando, los resultados son:

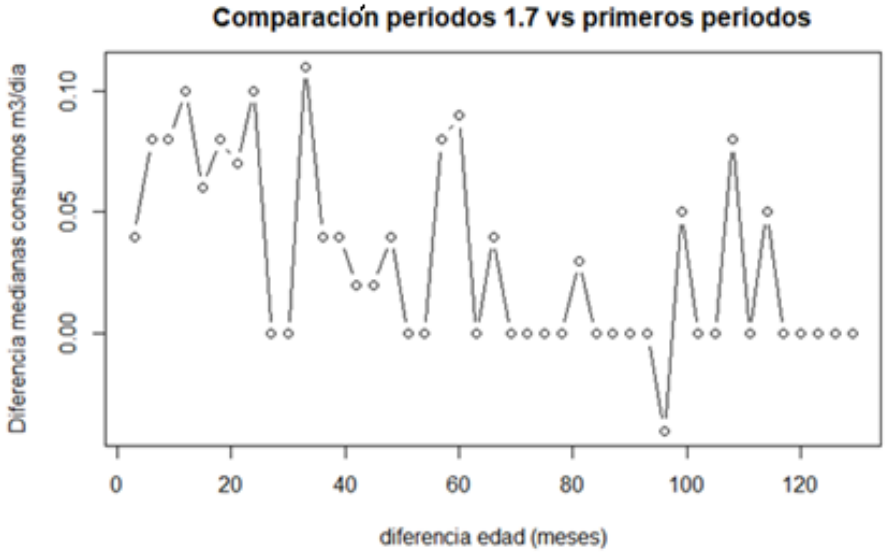
1.7 vs	Dato antiguo	Dato nuevo	p-valor	Ma - Mn	% diferen.
4.7	451	175	8.23e-05	0.04	29.5
7.7	430	101	1e-05	0.08	67.13
10.7	415	152	0	0.08	64.49
1.8	394	111	0	0.1	104.4
4.8	385	213	0	0.06	47.9
7.8	386	92	1.3e-06	0.08	57.61
10.8	367	78	2.62e-05	0.07	49.41
1.9	348	94	0	0.1	89.74
4.9	336	529	0.2581686	0	0
7.9	344	762	0.1051511	0	0
10.9	329	109	0	0.11	91.49
1.10	326	288	0.005895	0.04	24.35
4.10	321	283	0.000272	0.04	28.58
7.10	315	260	0.0341024	0.02	8.7
10.10	306	266	0.0432125	0.02	8.18
1.11	308	104	0.0186723	0.04	19.32
4.11	301	757	0.1353414	0	0
7.11	18	299	Fdd	0	Fdd
10.11	293	85	0.0001709	0.08	50.07
1.12	297	136	2e-07	0.09	61.8
4.12	279	557	0.0588285	0	0
7.12	278	220	0.0010487	0.04	25.71
10.12	266	401	0.8105872	0	0

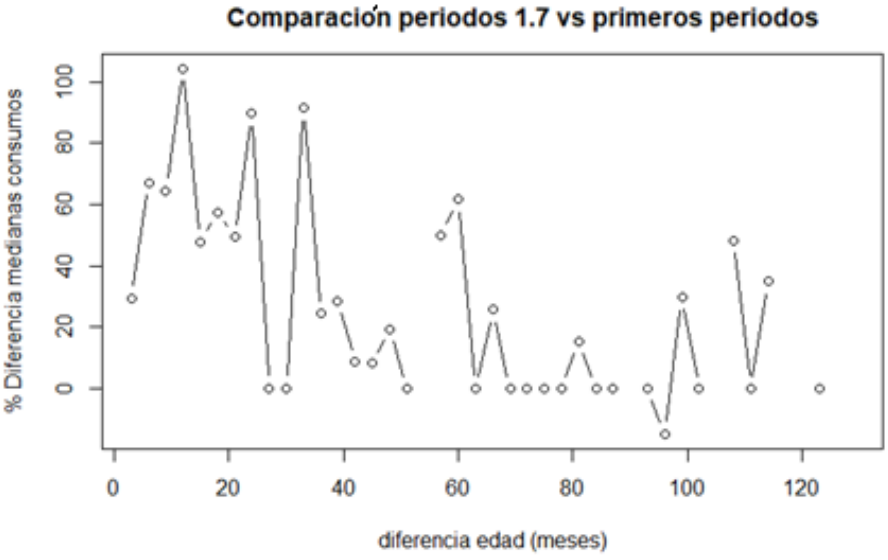
1.7 vs	Dato antiguo	Dato nuevo	p-valor	Ma - Mn	% diferen.
1.13	262	442	0.3233989	0	0
4.13	262	318	0.3551682	0	0
7.13	259	222	0.5777339	0	0
10.13	256	436	0.0018614	0.03	15.34
1.14	259	230	0.1963539	0	0
4.14	247	477	0.1544087	0	0
7.14	29	251	Fdd	0	Fdd
10.14	246	216	0.3044039	0	0
1.15	245	536	0.0067499	-0.04	-14.76
4.15	230	69	0.0089678	0.05	29.77
7.15	230	385	0.839234	0	0
10.15	29	221	Fdd	0	Fdd
1.16	224	71	0.0002886	0.08	48.34
4.16	225	576	0.6477754	0	0
7.16	224	141	0.0007973	0.05	34.93
10.16	29	210	Fdd	0	Fdd
1.17	26	206	Fdd	0	Fdd
4.17	159	158	0.630033	0	0
7.17	5	148	Fdd	0	Fdd
10.17	10	127	Fdd	0	Fdd

Si vamos por columnas de izquierda a derecha:

- Fecha contador instalado vs fecha contador instalado posteriormente (los de primeros consumos)
- Cantidad de datos de los sucesivos periodos de 1.7
- Cantidad de datos de las sucesivas muestras de primeras lecturas.
- P-valor de la prueba U de Mann-Whitney
- Diferencia de medianas (mediana del contador de más edad – mediana contador primeros consumos)
- % en la diferencia de medianas

Para poder observar estos datos mediante gráficas, en la primera de ellas se expone la diferencia de medianas entre las muestras y en la segunda el porcentaje de diferencia.

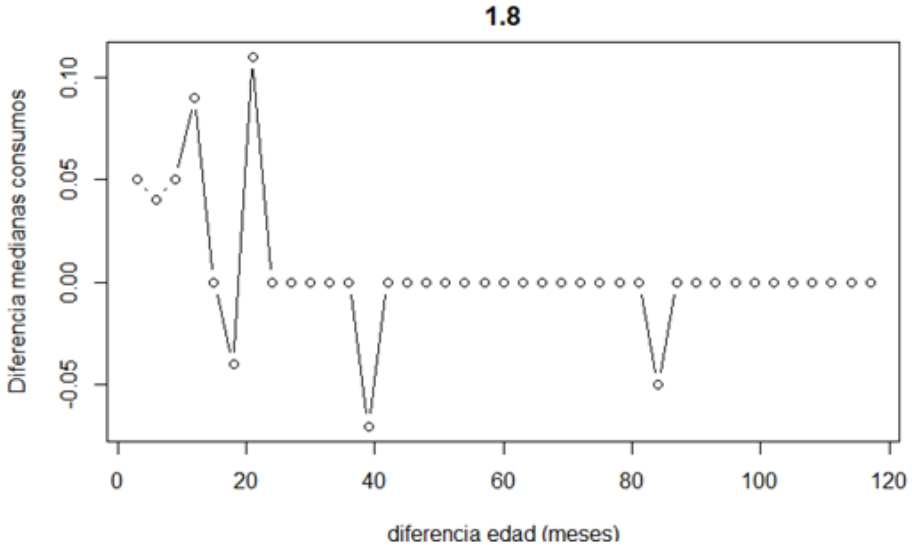




3.6 Resultados y conclusiones

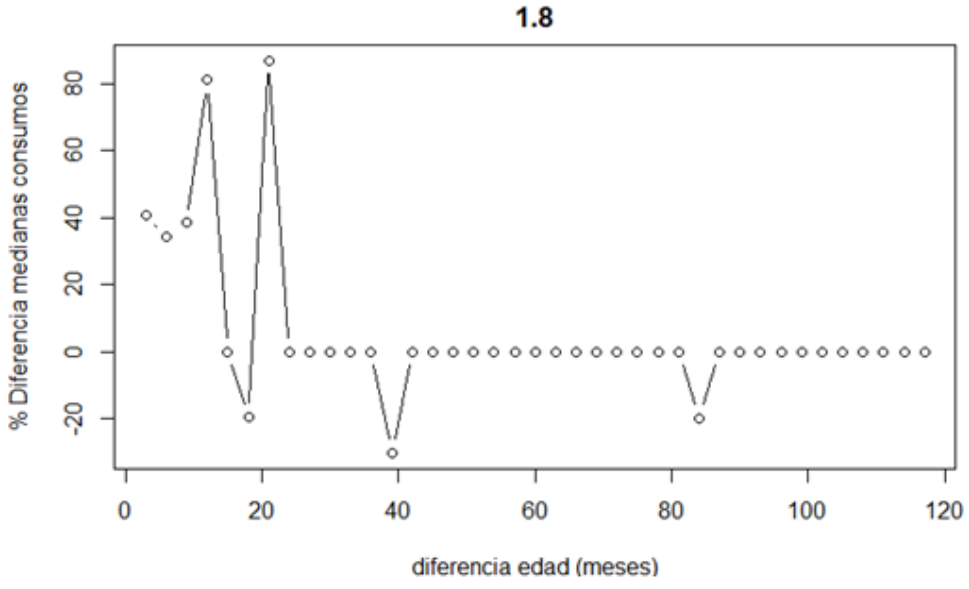
Aplicando el método expuesto para las muestras de contadores de distinta edad, podemos finalmente graficar la diferencia de medianas y su porcentaje de incremento o disminución.

Así, por ejemplo, podemos observar varios casos.



En esta gráfica observamos la diferencia de medianas de los sucesivos periodos de los contadores 1.8 respecto a lo marcado por recién instalados. Vemos que, en un gran número de periodos, no se encuentra diferencia entre lo marcado por las muestras de 1.8 y los nuevos. En los primeros periodos si se observa diferencia, llegando a alcanzar más de 0.10 m3/día, que es una cantidad bastante importante. Seguramente cuando las muestras difieren tanto, no se deba a una diferencia exclusiva en la degradación, sino que hay más diferencias en cuanto a las características de las muestras.

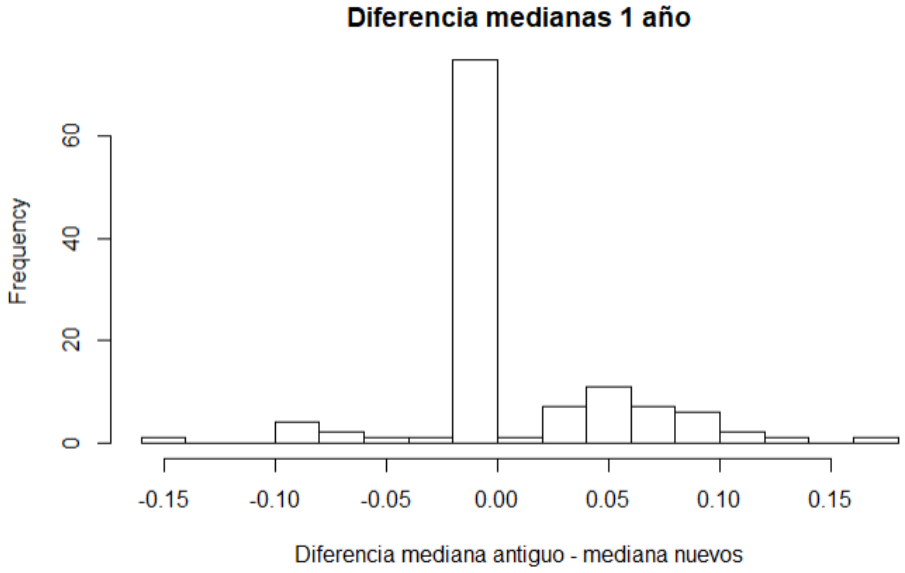
Obteniendo esta diferencia de medianas en porcentaje:



Donde igualmente apreciamos que hay una diferencia notable entre algunas muestras debidas, seguramente a factores externos y en una gran mayoría no se observa diferencia.

En el Anexo I se han expuesto estas y otras gráficas obtenidas en la aplicación de este primer método sobre datos de consumo trimestrales.

Al tener tantas comparaciones, lo más cómodo es resumir las diferencias de medianas que se encuentran entre las muestras mediante histogramas.

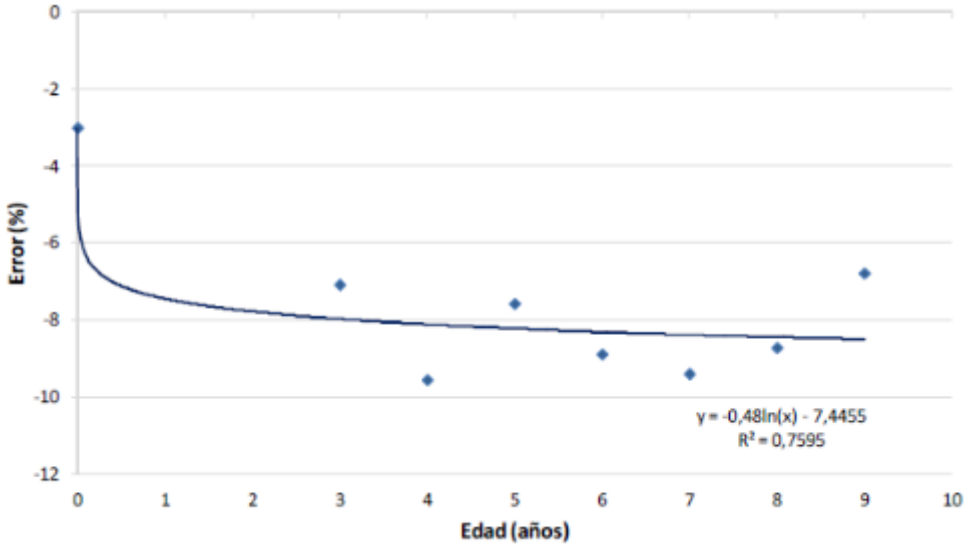


Aquí podemos ver un histograma para la diferencia de medianas entre muestras de contadores nuevos, y otras muestras de contadores que tenían 1 año de vida utilizando la regla de Freedman-Diaconis para discretizar los datos.

Observamos que la gran mayoría ha caído en la barra central, y la mayor parte de esas medidas son nulas, es decir, que en los contrastes de hipótesis entre contadores recién instalados y otros con una edad de 1 año, la gran mayoría de veces se acepta la hipótesis nula de medianas iguales, con lo que no se puede decir que las muestras difieran, por lo tanto, suponemos que no hay diferencia entre lo que marca un contador recién instalado y otro que lleva 1 año.

Igualmente se dispone de diferentes histogramas en el Anexo I para diferencias de edad mayores, y observándolos, vemos que casi la totalidad de histogramas coincide con lo expuesto en el anterior, es decir, no se encuentra diferencia entre lo marcado por contadores recién instalados y otros de más edad.

Los datos utilizados en el ejemplo de implementación de este primer método corresponden a un tipo de contador denominado “m1”. Gracias al estudio por degradación [2] en bancos de prueba llevado dentro de la empresa FACSA, conocemos la curva de error referente a la edad de este contador.



Analizando esta curva de error, en principio nos esperaríamos unos resultados en la implementación de nuestro método parecidos; en cambio, como hemos podido observar, no vemos este patrón definido, ya que no se observa este aumento progresivo en la degradación. Además, cabe destacar que normalmente los contadores antiguos no difieren en lo marcado los recién instalados.

Seguramente necesitemos mejorar la comprensión de los datos de partida, obteniendo más características de su procedencia como puede ser ubicación, tipo de suministro, tipo de instalación, etc. e intentando homogenizar las muestras a comparar para poder mejorar el análisis.

Por otra parte, al realizar los histogramas de estas diferencias de medianas, sí que observamos que una gran mayoría de veces, el contador de más edad ha marcado igual que los nuevos, es decir, que no hay diferencia de medianas entre los contadores antiguos y los nuevos. Como vemos en la curva de degradación del contador, se produce un aumento bastante grande al principio, más o menos en los 3 primeros meses, y después el error aumenta de forma progresiva y suave. Esto sí coincide con los resultados obtenidos en los histogramas, ya que necesitaríamos una diferencia mayor en la degradación para poder percibir cambios entre las muestras.

Esta diferencia significativa entre el consumo marcado por los contadores de más edad y los nuevos sería necesaria ya que hay que tener en cuenta que estamos utilizando un estadístico no paramétrico. Este hecho hace que, al comparar dos muestras, se requiera una diferencia más notable que si estuviéramos utilizando un estadístico paramétrico; por ello, si la diferencia entre las dos muestras es pequeña, con este estadístico sería más difícil de detectar.

En la curva de error del contador sí existe una diferencia significativa, pero se produce al principio, casi recién instalado. En nuestro caso, al trabajar con consumos trimestrales, el

primer dato que obtenemos de un contador es a partir de los 3 meses, por lo que no podríamos notar este salto en la degradación trabajando sobre consumos trimestrales.

Capítulo 4

Segundo método: comparación entre la degradación en diferentes edades

4.1 Introducción

El objetivo de este método no es estudiar la degradación en sí, sino estudiar cómo varía en una edad respecto a otra.

Este estudio también se podría realizar con el anterior método, pero el objetivo del proyecto es aportar más herramientas para mejorar el conocimiento del contador a través de datos de consumo, es por ello que se ha implementado un método diferente.

Aquí también se utilizará el contraste de hipótesis como base, aunque se realizará entre datos de un mismo contador tomando periodos diferentes.

4.2 Primeros análisis

Utilizaremos igualmente datos de consumo trimestral para la explicación de este método.

Si tomamos los contadores 1.7 (instalados en enero del 2007) sabemos que tenemos datos de consumo en:

Mes/año	*1/7*	*2/7*	*3/7*	*4/7*	*5/7*	*6/7*	*7/7*	*…*	*10/7*	*…*	*1/8*	*…*
Consumo	*Inst.*	-	-	*X*	-	-	*X*	*…*	*X*	*…*	*X*	*…*
Periodo				*1*			*2*	*…*	*3*	*…*	*4*	*…*

Lo que podemos realizar ahora es una comparación entre lo consumido en diferentes periodos. Por ejemplo, podemos ver el porcentaje de incremento o descenso entre los periodos 2 y 3, que se correspondería a una diferencia de 3 meses entre ambas muestras. También podemos tomar los periodos 5 y 9, que se correspondería a una diferencia de 1 año y para mismos trimestres en años sucesivos, en donde el contador tendría una edad de 1 año y 3 meses en el 5º periodo y de 2 años y 3 meses en el 9º periodo.

Si comparamos las muestras de los periodos 5 y 9 para los 1.7, mediante U de Mann-Whitney podemos verificar si difieren o no en mediana.

	Tamaño muestral 5º periodo	**Tamaño muestral 9º periodo**	**p-valor U Mann-Whitney**	**Diferencia de medianas M5 – M9**	**% Diferencia de medianas**
Contadores 1.7	394	348	0.3513	0	0

En este caso observamos que no se puede decir que las muestras difieran en mediana.

Esta comparación entre los periodos 5 y 9 de los contadores 1.7 nos informa sobre si el consumo ha variado de un año a otro, pero esta variación no debe ser entendida en términos exclusivos de degradación, ya que existirán factores externos que puedan modificarla. Para poder sacar conclusiones sobre la degradación, lo que realizamos es otra comparación de muestras que hayan sido tomadas en el mismo espacio de tiempo con contadores de otra edad.

Así, si por ejemplo tomamos el periodo 1 y 5 de los contadores 1.8, observamos que se corresponden en tiempo a las muestras 5 y 9 de los contadores 1.7, en donde en el periodo 1 los contadores 1.8 tendrían una edad de 3 meses y en el periodo 5 una edad de 1 año y 3 meses.

Mes instalación/mes lectura	*1/8*	*4/8*	*7/8*	*10/8*	*1/9*	*4/10*
1.7	-	***5º periodo***	*6º periodo*	*7º periodo*	*8º periodo*	***9º periodo***
1.8	*Instalación*	***1º periodo***	*2º periodo*	*3º periodo*	*4º periodo*	***5º periodo***

Si realizamos la misma comparación para los periodos 1 y 5 de los 1.8 obtenemos

	Tamaño muestral 1º periodo	**Tamaño muestral 5º periodo**	**p-valor U Mann-Whitney**	**Diferencia de medianas M1 – M5**	**% Diferencia de medianas**
Contadores 1.8	111	111	0.00057	0.012	4,3

Al ser el p-valor menor del nivel de significación tomado (0,05), obtenemos la diferencia de medianas (mediana muestra periodo 1 – mediana muestra periodo 5) y el porcentaje de esta diferencia.

Si hallamos la diferencia entre el "% diferencia de medianas" de las muestras 1.7 y 1.8, se obtiene que los contadores de la muestra 1.8 se han degradado un 4,3% más que los contadores de la muestra 1.7.

Si lo medimos en edad, significa que entre la edad de 3 meses (1º periodo 1.8) hasta 1 años y 3 meses (5º periodo 1.8) se ha producido una degradación un 4,3 % mayor que entre la edad de 1 año y 3 meses (5º periodo 1.7) hasta 2 años y 3 meses (9º periodo 1.7).

Para verlo de manera más clara, podemos imaginar contadores ideales que no tienen ningún tipo de degradación, entonces, si las dos muestras están expuestas a las mismas condiciones externas y tienen las mismas características, la diferencia entre los periodos 5 y 9 de los contadores 1.7 debería ser la misma que la diferencia entre los periodos 1 y 5 de los contadores 1.8. Ahora bien, si esta diferencia no es la misma, deberá de existir un factor que no sea el mismo entre ambos. Ese factor es la degradación.

Por lo tanto, lo que estamos realizando con este procedimiento no es medir la degradación en sí, es comparar si existe una diferencia de degradación entre contadores de diferentes edades a lo largo de un año.

4.3 Comparación de consumos: U de Mann-Whitney

El contraste de hipótesis tiene lugar cuando comparamos en unos mismos contadores sus periodos (en el caso anterior 5º y 9º para los 1.7 y 1º y 5º para los 1.8).

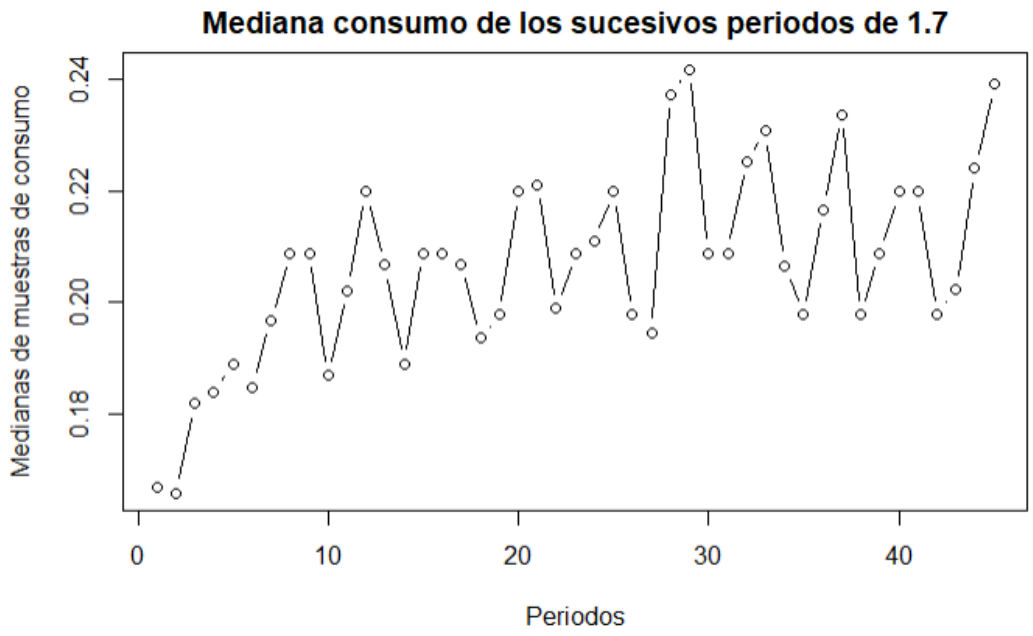
Se realiza el mismo contraste de U de Mann-Whitney del primer método, tomando el mismo nivel de significación $\alpha = 0,05$.

Con ello verificamos si los consumos marcados por un conjunto de contadores en diferentes periodos difieren entre sí o no en mediana.

4.4 Aplicación sobre datos de consumo trimestral

Al igual que en el anterior método, vamos a realizar una explicación más detallada para familiarizarnos mejor con este proceso.

En primer lugar, podemos observar que los datos tienen estacionalidad:



Es decir, hay periodos dentro del año donde se consume más y otros menos, pero esta estacionalidad no es la misma para los contadores instalados en diferentes meses o años, es por eso que en la comparación para una misma muestra de contadores en dos periodos diferentes hemos tomado periodos que pertenezcan a la misma época, pero de diferentes años.

Realizando ahora la comparación entre los periodos 5 y 9 de las muestras de contadores instalados desde enero de 2007 a diciembre de 2007, tenemos:

	Nº Datos 5º periodo del 2007	Nº Datos 9º periodo del 2007	p-valor	M5º - M9º	% diferencia
Enero 1.7	394	348	0.353	0	0
Febrero 2.7	346	305	0.631	0	0
Marzo 3.7	242	193	0.373	0	0
Abril 4.7	160	143	0.281	0	0

	Nº Datos 5º periodo del 2007	Nº Datos 9º periodo del 2007	p-valor	M5º - M9º	% diferencia
Mayo 5.7	204	179	0.933	0	0
Junio 6.7	15	15	Fdd	0	0
Julio 7.7	102	78	0.713	0	0
Agosto 8.7	74	67	0.885	0	0
Septiembre 9.7	117	106	0.789	0	0
Octubre 10.7	139	125	0.626	0	0
Noviembre 11.7	191	173	0.664	0	0
Diciembre 12.7	215	206	0.417	0	0

Podemos observar por columna, de izquierda a derecha:

- Cantidad de datos del 5º periodo de contadores instalados en 2007 según mes
- Cantidad de datos del 9º periodo de contadores instalados en 2007 según mes
- P-valor según U de Mann-Whitney por contraste de hipótesis en mediana de las dos anteriores muestras
- Diferencia de medianas: periodo 5 – periodo 9
- % diferencia entre las medianas

Obtenemos una nueva tabla que coincida en el tiempo con la anterior pero que difiera en edad.

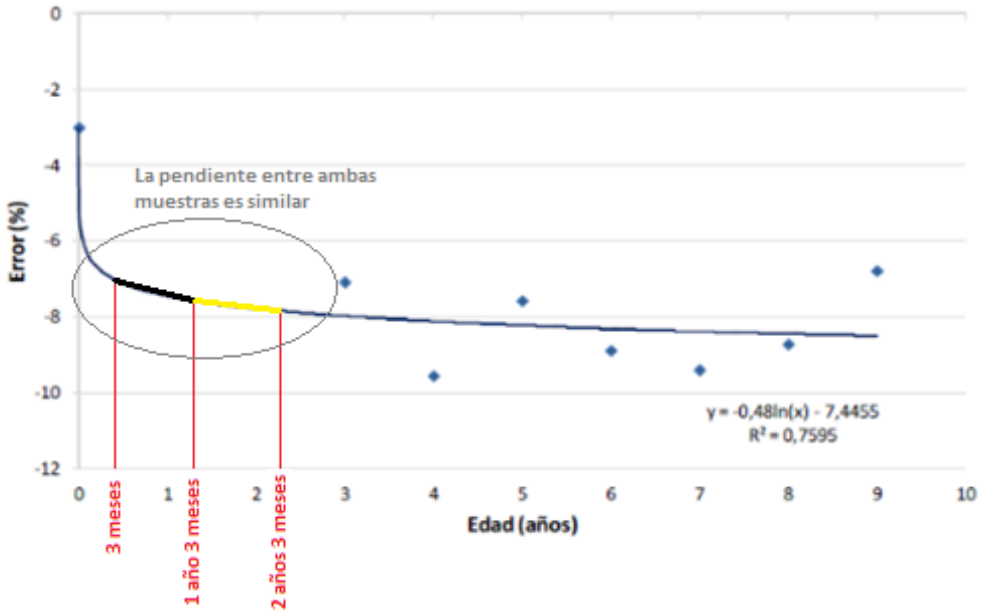
Si tomamos contadores instalados en 2008, sus periodos 1 y 5 coinciden en tiempo con los anteriores y difieren en edad, que es lo que buscamos.

	Nº Datos 1º periodo del 2008	Nº datos 5º periodo del 2008	p-valor	M1º - M5º	% diferencia
Enero 1.8	111	111	0.005	0.08	44.7
Febrero 2.8	180	169	0.141	0	0
Marzo 3.8	213	192	0.004	0.05	36.6
Abril 4.8	213	193	0.345	0	0
Mayo 5.8	189	170	0.432	0	0
Junio 6.8	180	106	0.215	0	0
Julio 7.8	92	83	0.998	0	0

	Nº Datos 1º periodo del 2008	Nº datos 5º periodo del 2008	p-valor	M1º - M5º	% diferencia
Agosto 8.8	154	150	0.007	0.05	36.7
Septiembre 9.8	16	22	Fdd	0	0
Octubre 10.8	78	75	0.265	0	0
Noviembre 11.8	124	117	0.301	0	0
Diciembre 12.8	88	82	0.825	0	0

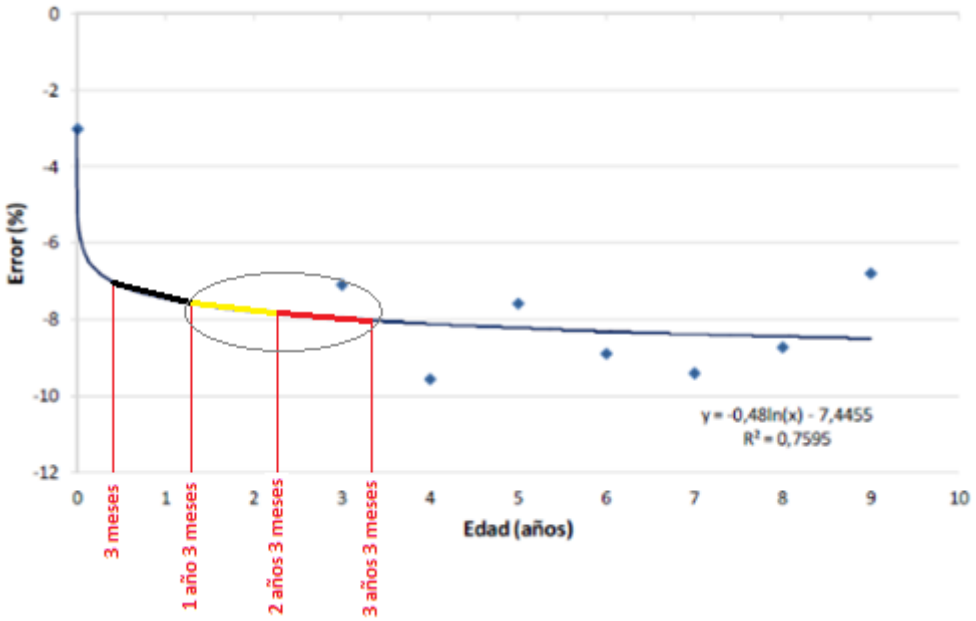
Lo que nos falta es simplemente obtener la diferencia entre los porcentajes de ambas tablas. Vemos en este caso, que la mayoría de las veces no existe diferencia entre la degradación en estos periodos de tiempo, por consiguiente, tampoco entre la degradación comprendida entre las edades de [3 meses,1 año y 3 meses] y [1 año y 3 meses, 2 años y 3 meses].

Para verlo de forma gráfica, en la curva de degradación obtenida en bancos de prueba marcamos las pendientes en la degradación entre las edades de estudio. Vemos que no difieren mucho, que es acorde a los resultados obtenidos en nuestro procedimiento.



Como continuación, tomaríamos ahora el 5º y 9º periodo para los 1.8 y el 1º y 5º periodo para los 1.9, obtenemos nuevamente la diferencia de porcentajes y seguimos repitiendo el método. Con ello estaríamos calculando la diferencia del porcentaje en la degradación que se produce entre [3 meses,1 año y 3 meses] y [1 año y 3 meses, 2 años y 3 meses] para diferentes muestras.

En el siguiente paso, tomaríamos de los contadores 1.7 el 9º y 13º periodo y del 1.8 el 5º y 9º. Con lo que obtendríamos la diferencia del porcentaje en la degradación entre [1 año y 3 meses,2 años y 3 meses] y [2 años y 3 meses, 3 años y 3 meses] como vemos en la gráfica.

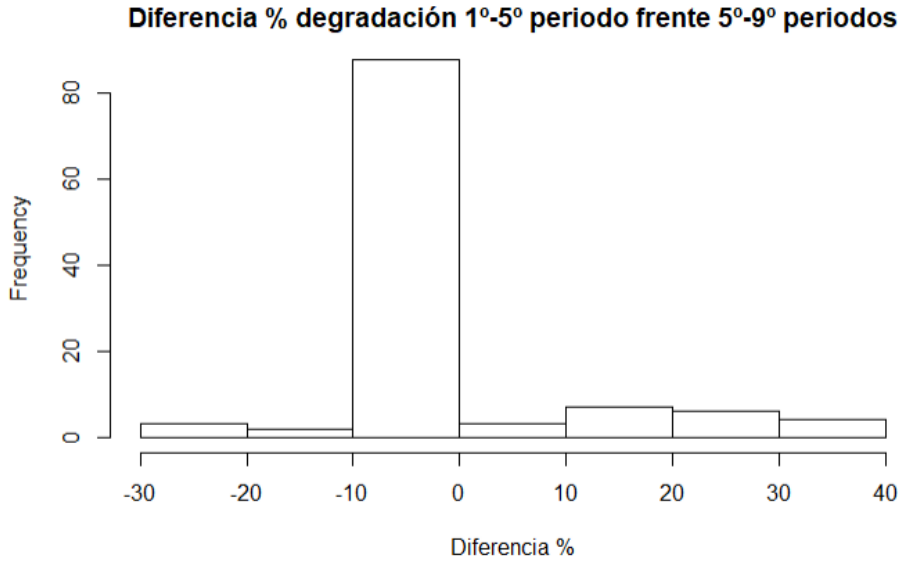


Procediendo de la misma manera seguiríamos comparando el porcentaje de degradación para las sucesivas edades.

4.5 Resultados y conclusiones

En la aplicación de este segundo método para datos de consumo trimestral, se han expuesto en el anexo II histogramas clasificados por los periodos entre los cuales se realiza la comparación de las muestras.

Si tomamos el siguiente ejemplo:



Observamos la diferencia en tanto por ciento entre lo que ha aumentado o disminuido lo marcado por los contadores entre su 1º y 5º periodo respecto a lo que ha aumentado o disminuido entre su 5º y 9º periodo.

Tomando dominio temporal:

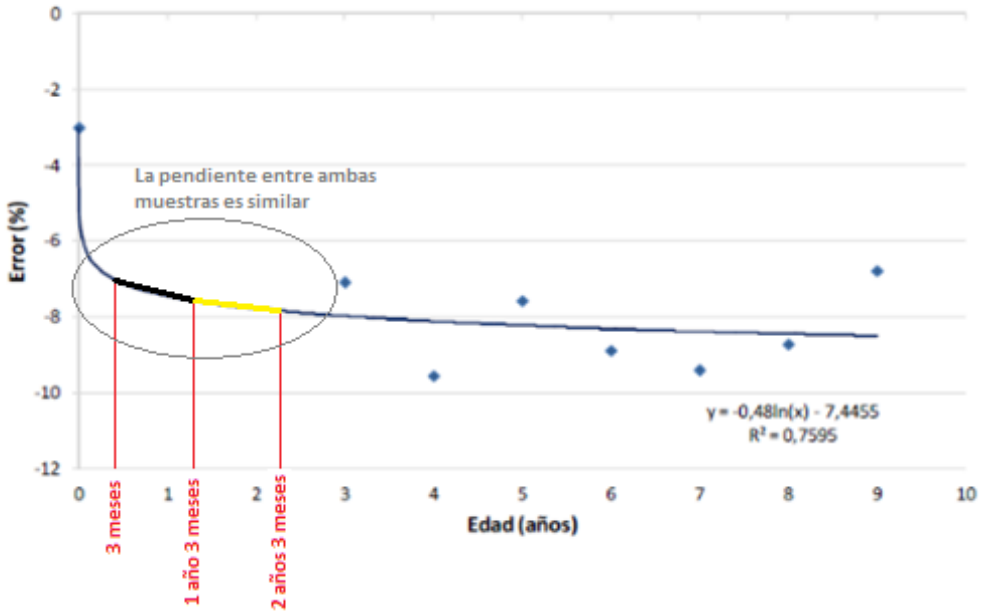
1º periodo = 3 meses

5º periodo = 1 año y 3 meses

9º periodo = 2 años y 3 meses

Por lo que finalmente vemos que implica un estudio entre lo que se degrada el contador entre 3 meses y 1 año y 3 meses respecto a lo que se degrada entre 1 año y 3 meses hasta 2 años y 3 meses.

De manera gráfica:

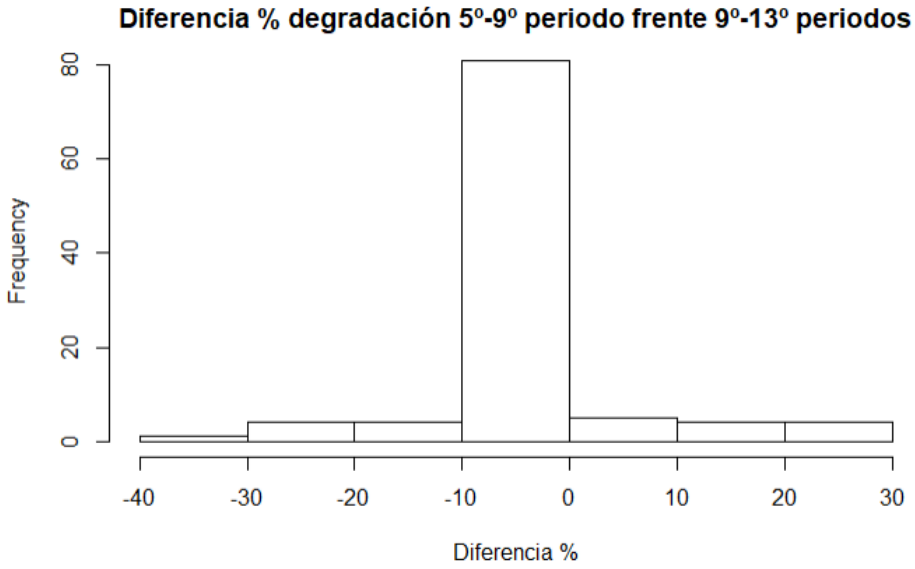


De lo que nos informa este primer histograma es sobre la diferencia entre las pendientes marcadas en negro y amarillo sobe la curva de error del contador obtenida en bancos de prueba.

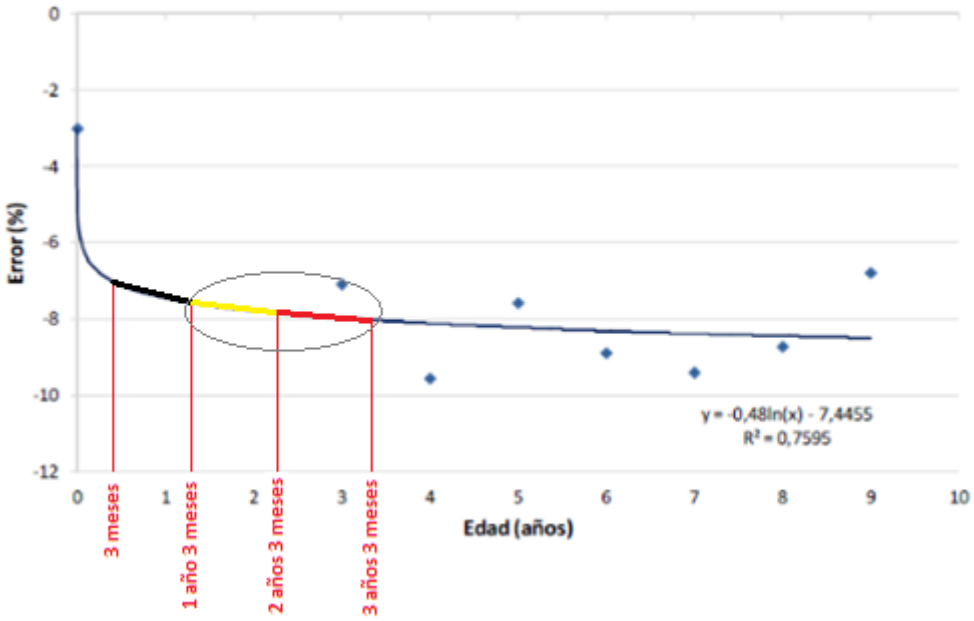
Observamos que el histograma nos informa que en su gran mayoría no se ha detectado o no se puede concluir que las muestras difieran, lo que coincide con lo visto en la gráfica sobre el porcentaje de error del contador obtenido en bancos de prueba.

Tomando más histogramas de este anexo, vemos que en la gran mayoría de muestras no difiere lo marcado por los contadores entre sus diferentes periodos.

Si seguimos con el siguiente paso vemos:



Y realizando las pendientes sobre la curva de error del contador:



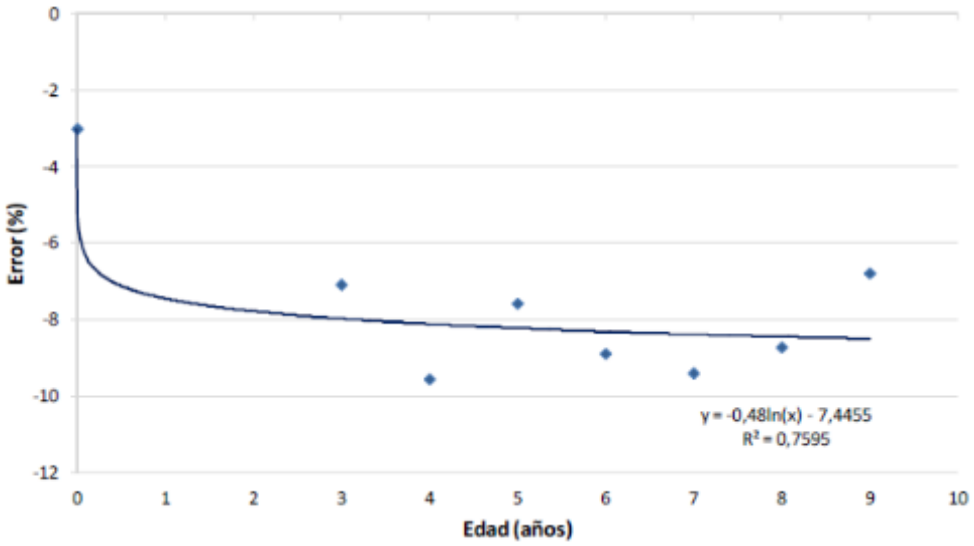
Observamos que las pendientes son similares y es lo obtenido en el histograma.

Como las lecturas que estamos utilizando son trimestrales, estos periodos se corresponden a diferentes edades del contador (1º periodo: 3 meses; 5º periodo: 1 año y 3 meses; 9º periodo: 2 años y 3 meses), como ya habíamos visto.

Aquí vemos gráficamente lo que se ha comparado en este primer caso, las pendientes en la degradación entre las edades marcadas.

A lo largo del Anexo II encontramos estas comparaciones realizadas entre diferentes periodos, que se corresponden a la diferencia observada en la degradación para distintas edades del contador y en las cuales, la gran mayoría nos informa que no hay diferencias entre pendientes a lo largo de la vida del contador.

Si volvemos a mirar la curva de degradación del contador, la degradación que más difiere con respecto al resto se da al principio, y el resto parece tener una pendiente en su degradación parecida.



Esto se corresponde a los resultados obtenidos, ya que la gran mayoría de veces no notamos una diferencia entre las muestras, lo que hace que la degradación sea parecida a lo largo de la edad.

Una vez más, observamos que el descenso más marcado en la curva de degradación se corresponde al principio en la vida del contador, y, como nos pasó en el anterior método, al trabajar con consumos de medida trimestral, no notamos este marcado descenso en la curva.

Las comparaciones se hicieron entre periodos que comprendían 1 año de diferencia, para no estar expuestos a la comparación de datos con diferentes estacionalidades; sin embargo, si tuviéramos las suficientes características de los datos de la muestra, quizás pudiéramos seleccionar aquellos con similares estacionalidades, con lo que podríamos realizar comparaciones más cercanas en tiempo y obtener mayor información.

Capítulo 5

Conclusiones y trabajo futuro

A lo largo del proyecto, hemos podido comprobar la dificultad que entraña el estudio de la degradación de un contador. Y este ha sido uno de los principales desafíos, enfrentarse a una problemática conocida desde un contexto diferente.

En el análisis descriptivo de los datos vimos como la tendencia del consumo es, en general, a la baja, donde los factores externos influyen notablemente sobre esta tendencia, dándose por una mayor eficiencia de los electrodomésticos, una mayor concienciación, etc.

Para mejorar nuestro análisis gráfico hicimos uso de media móvil. Esta media móvil es una manera de ver la tendencia de forma más suave. Aquí también se podría derivar otro tipo de estudio:

- Si disponemos los consumos como una serie temporal y esta tendencia en vez de obtenerla por media móvil, la obtenemos por métodos de regresión, ya sea lineal, logarítmica o exponencial, estaríamos capacitados para predecir valores futuros de consumo, lo cual podría ser interesante a la hora de hacernos una idea de la facturación, del caudal residual que deberemos tratar o inclusive para el cálculo de nuevas redes de abastecimiento.

Mediante el método K-NN realizamos un estudio para comparar las curvas de consumo para diferentes muestras con contadores de distinta edad. Con este método lo que estudiamos fue si las curvas de consumo tenían patrones distintos para cada muestra.

En los resultados obtuvimos un mayor porcentaje de éxito para las curvas de consumo de mayor edad, pero este porcentaje tampoco fue muy elevado (menor del 40 %) y tampoco se observó un patrón de los aciertos en dependencia de la edad.

En la implementación de los diferentes métodos, se partía de un estudio sobre las curvas de degradación del contador con el que trabajábamos realizado en bancos de prueba. Era importante este estudio para poder comparar los resultados obtenidos en este proyecto.

Con la implementación del primer método, intentamos obtener un patrón en la degradación debida a la diferencia de edad del contador. Los resultados no fueron del todo satisfactorios en este aspecto, ya que no se consiguió un patrón, pero, observando los resultados globales para las distintas comparaciones, vimos que apenas obtuvimos diferencias significativas entre la degradación a diferentes edades. Este hecho implica que, teniendo en cuenta que utilizamos un estadístico no paramétrico, con lo que necesitamos una distinción mayor entre las muestras para rechazar la hipótesis nula de igualdad en medianas, sí se corresponde a la curva de degradación obtenida en bancos de prueba. En esta curva, inicialmente hay un fuerte aumento en la degradación, pero después este acrecentamiento se produce de forma suave. En nuestro estudio lo que obtuvimos es que, la gran mayoría de veces, no se podía observar diferencia entre lo marcado a distintas edades.

El segundo método se implementó para comparar la degradación a diferentes edades del contador y ver si había cambios en esta pendiente. Comparamos la degradación que sufría un contador entre sus 3 primeros meses de vida y el año, frente a la que sufría entre el primer y segundo año de vida. Después comparamos la degradación que sufría entre su primer y segundo año de vida frente a la que sufría entre su tercer y cuarto año activo. Realizando este proceso sucesivamente y viendo los resultados, no se observaba que la pendiente de la degradación cambiara a lo largo de la vida del contador.

Este resultado sí se corresponde a la curva obtenida en bancos de prueba para este tipo de contador, ya que, como dijimos antes, esta curva aumenta de forma suave, haciéndose difícil observar diferencia entre la pendiente marcada entre sucesivos años.

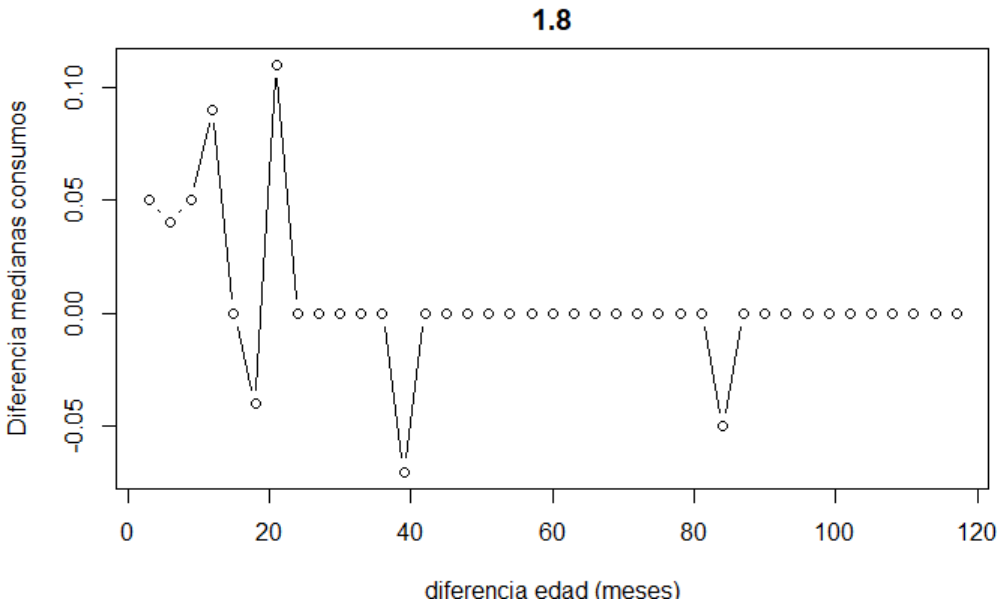
Los métodos aquí propuestos son un primer paso para el estudio de la degradación a partir de datos de consumo, si bien es cierto que están pensados para una implementación sobre datos donde dispongamos de todas las características posibles, los resultados obtenidos han sido razonables.

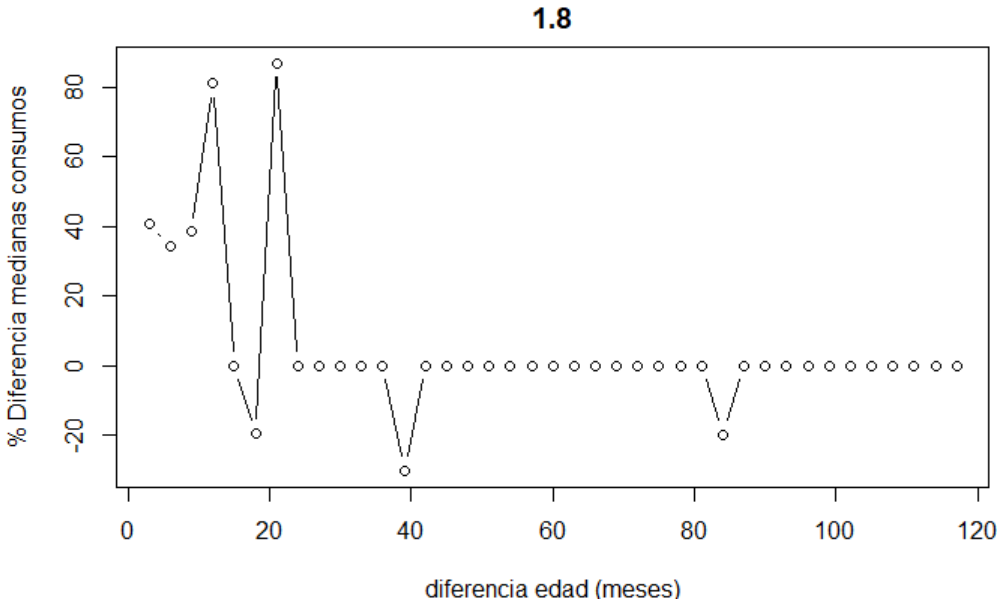
La continuación lógica de este estudio sería la que apuntamos anteriormente, implementar estos métodos bajo datos con todas las características posibles y en diferentes modelos de contador y concluir su eficacia, puntos de mejorar o incluso adaptarlos según la naturaleza del estudio.

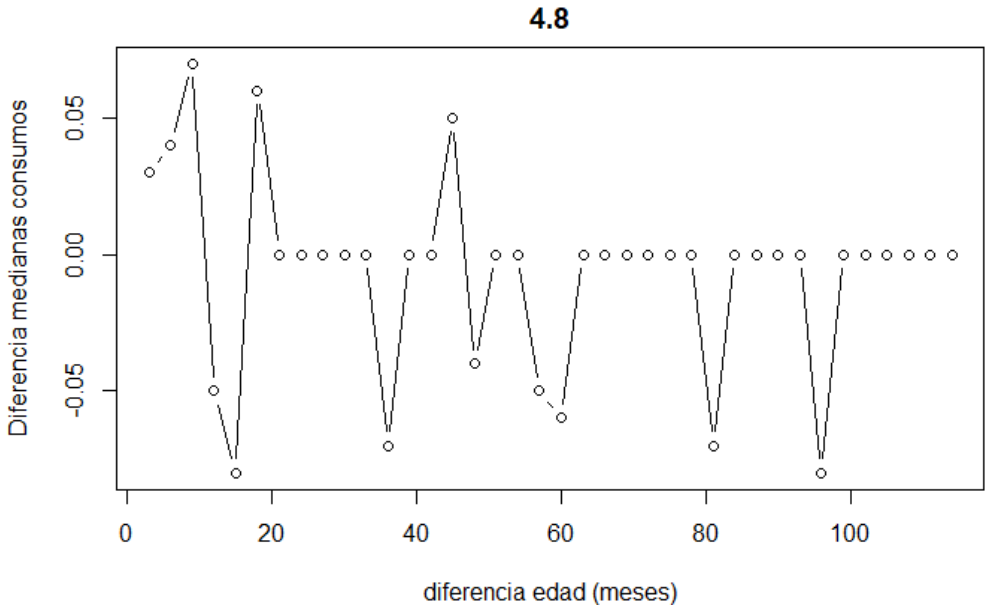
Anexos

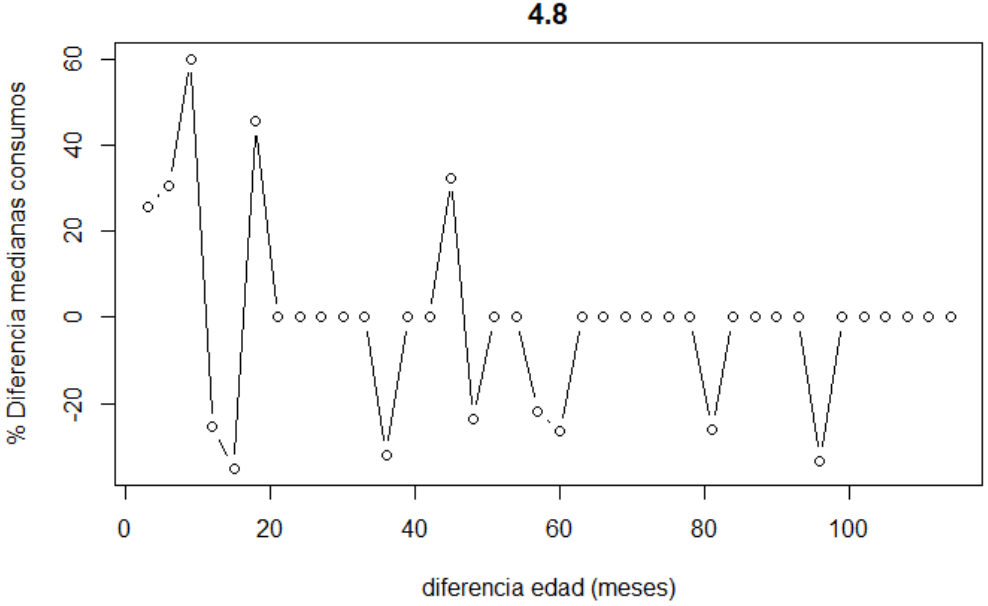
Anexo I

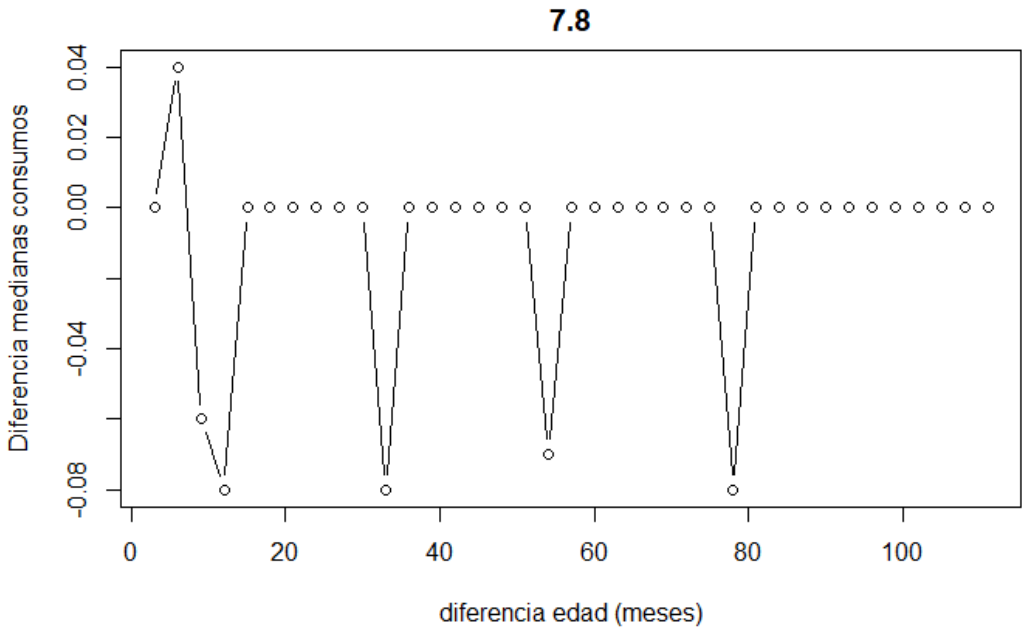
Primer método de estudio

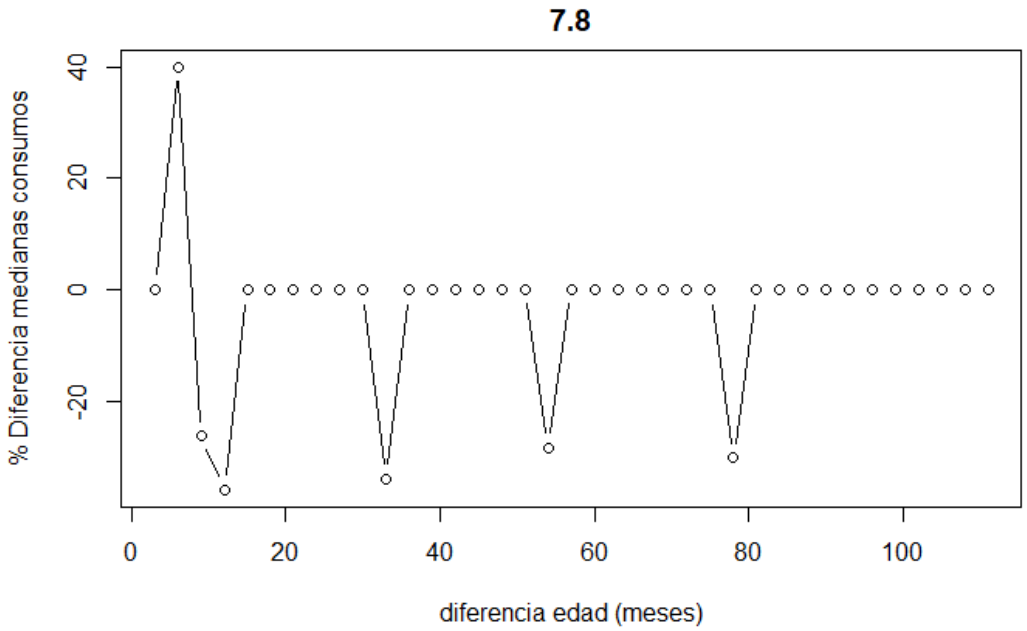

1.8
Diferencia medianas consumos
-0.05
0.00
0.05
0.10
0
20
40
60
80
100
120
diferencia edad (meses)

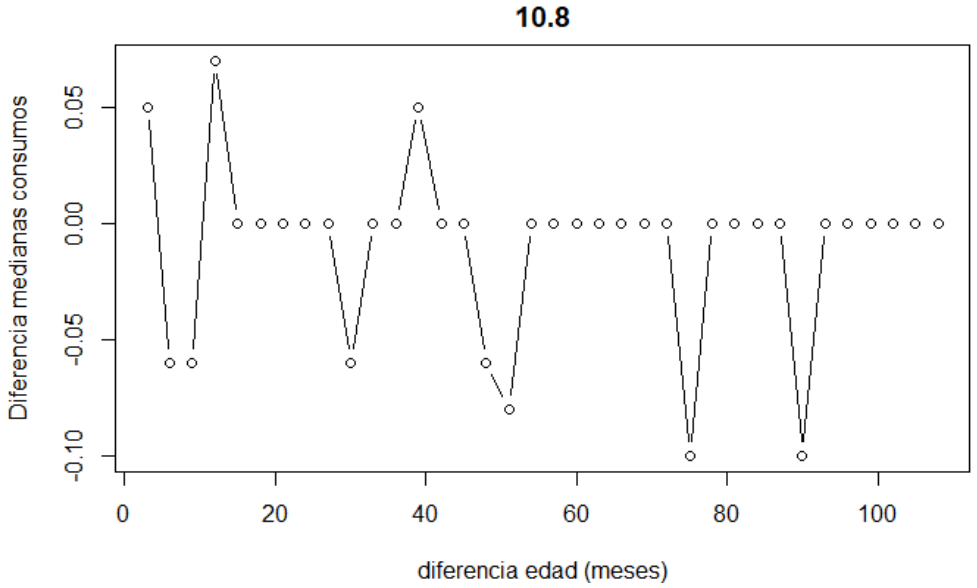

1.8
% Diferencia medianas consumos
-20
0
20
40
60
80
0
20
40
60
80
100
120
diferencia edad (meses)

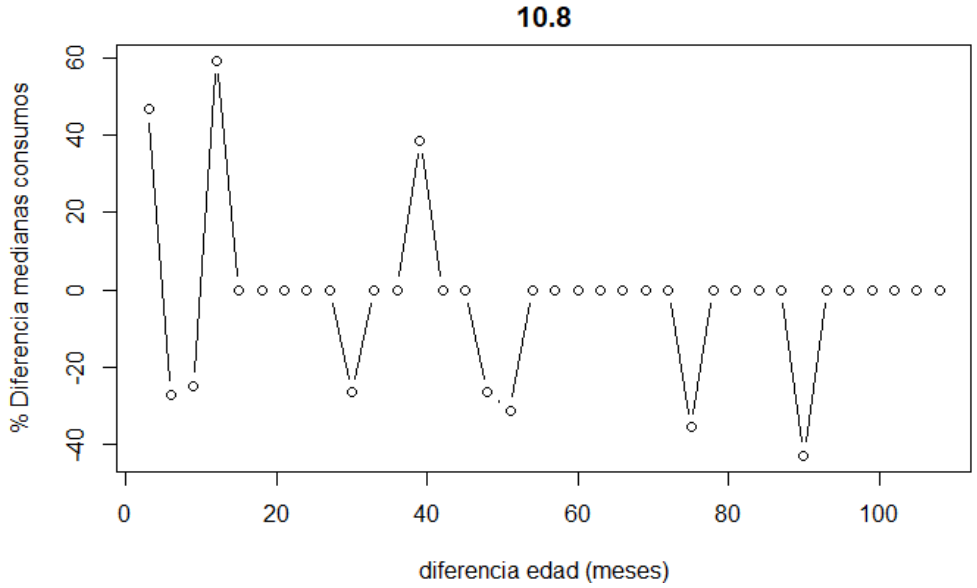

4.8
Diferencia medianas consumos
-0.05
0.00
0.05
0
20
40
60
80
100
diferencia edad (meses)

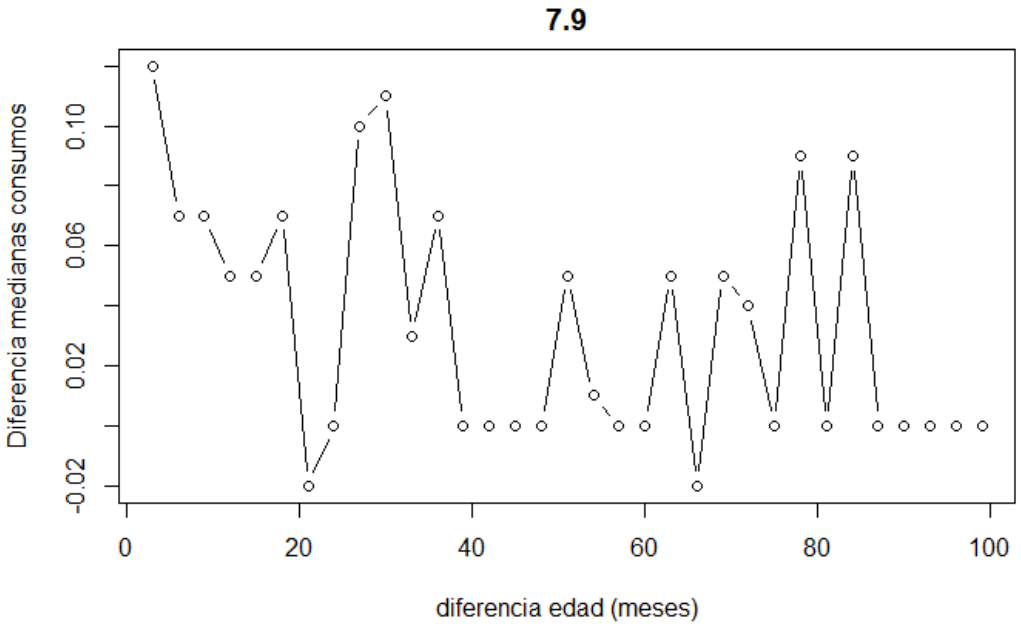

4.8
% Diferencia medianas consumos
-20
0
20
40
60
0
20
40
60
80
100
diferencia edad (meses)

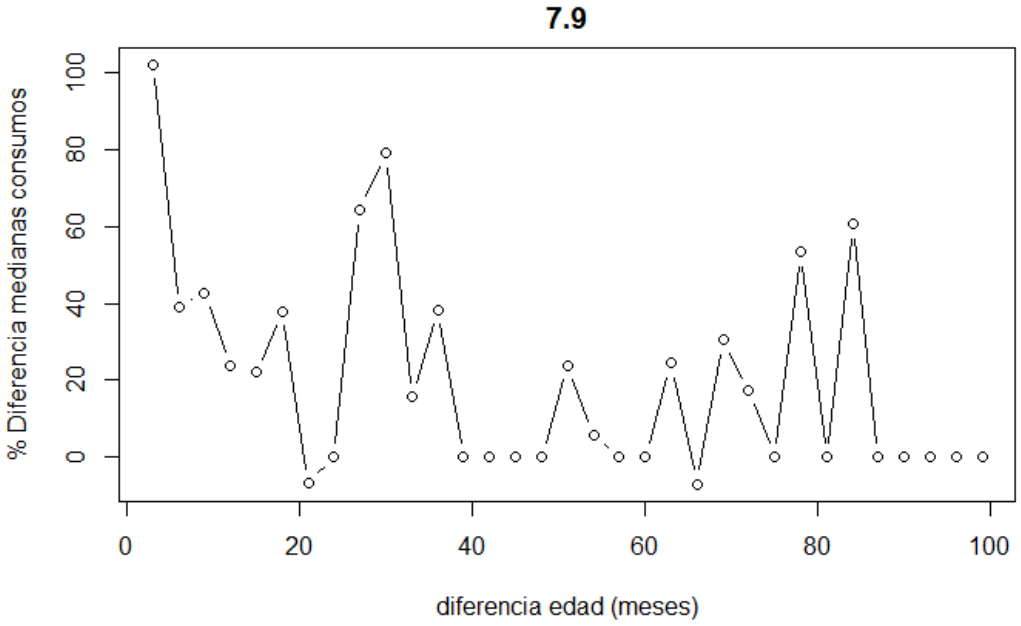

7.8
Diferencia medianas consumos
-0.08
-0.04
0.00
0.02
0.04
0
20
40
60
80
100
diferencia edad (meses)

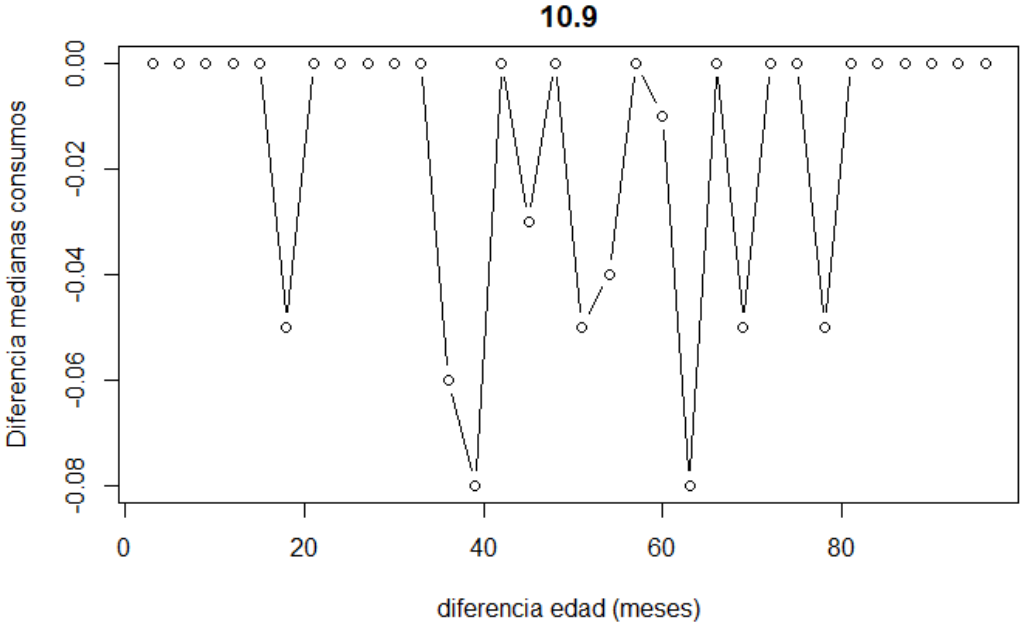

7.8
% Diferencia medianas consumos
-20
0
20
40
0
20
40
60
80
100
diferencia edad (meses)

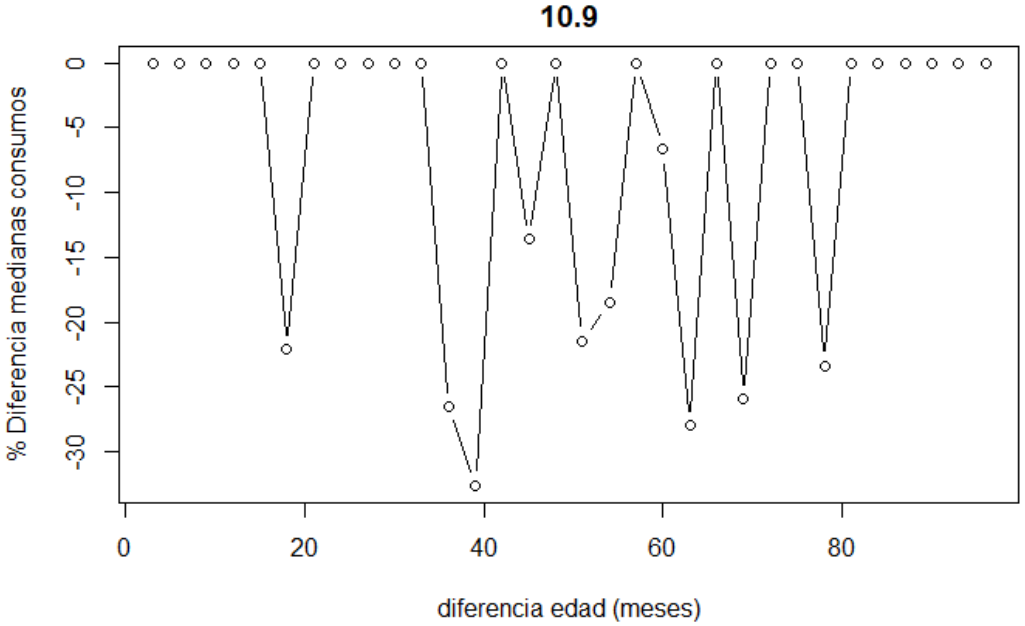

10.8
Diferencia medianas consumos
-0.10
-0.05
0.00
0.05
0
20
40
60
80
100
diferencia edad (meses)

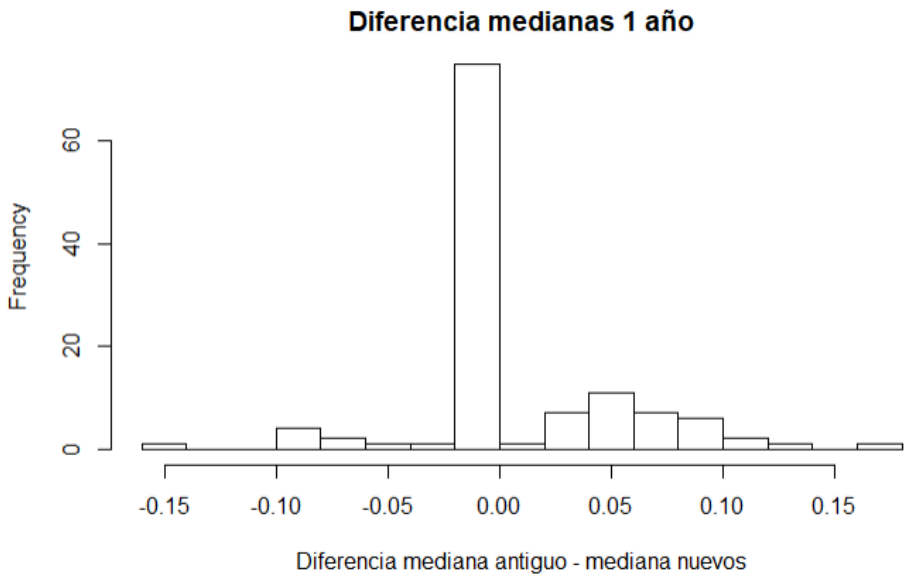

10.8
% Diferencia medianas consumos
-40
-20
0
20
40
60
0
20
40
60
80
100
diferencia edad (meses)

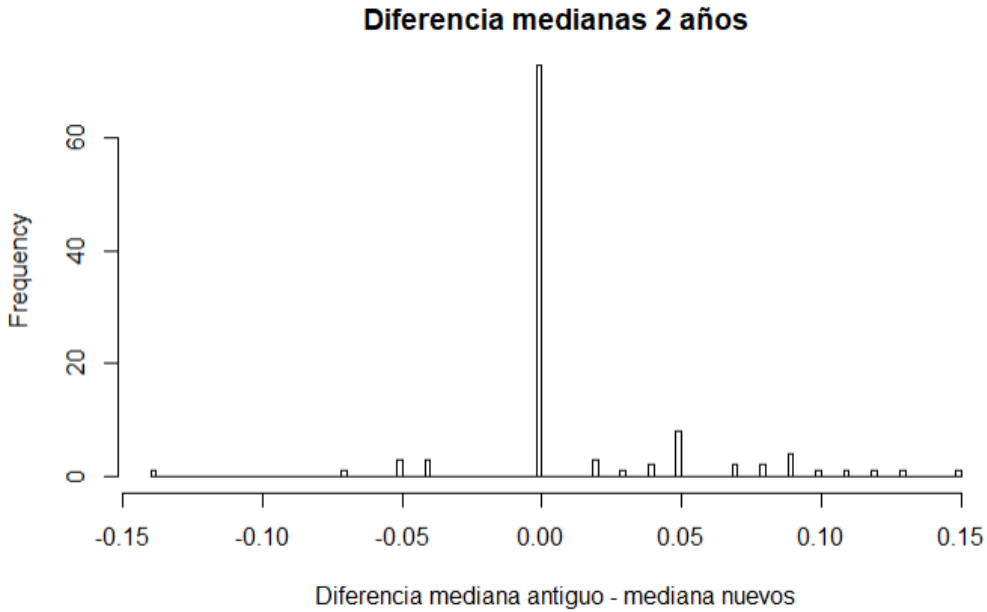

7.9
Diferencia medianas consumos
-0.02
0.02
0.06
0.10
0
20
40
60
80
100
diferencia edad (meses)

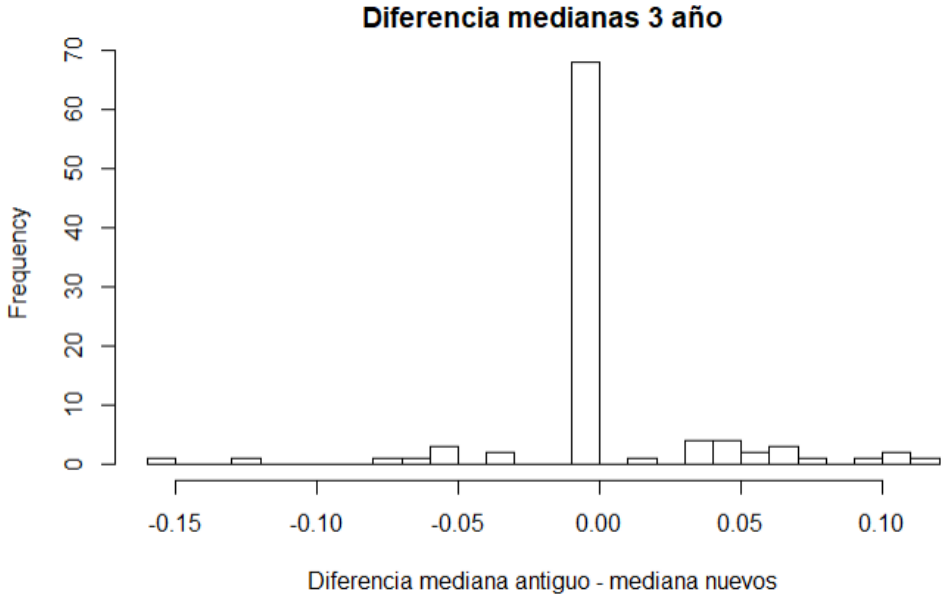

7.9
% Diferencia medianas consumos
0
20
40
60
80
100
0
20
40
60
80
100
diferencia edad (meses)

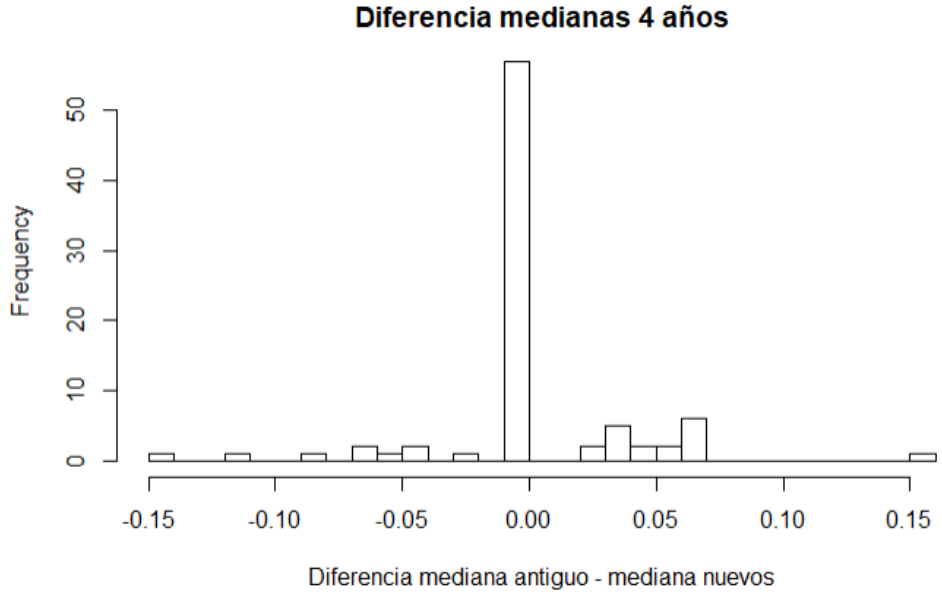

10.9
Diferencia medianas consumos
-0.08
-0.06
-0.04
-0.02
0.00
0
20
40
60
80
diferencia edad (meses)

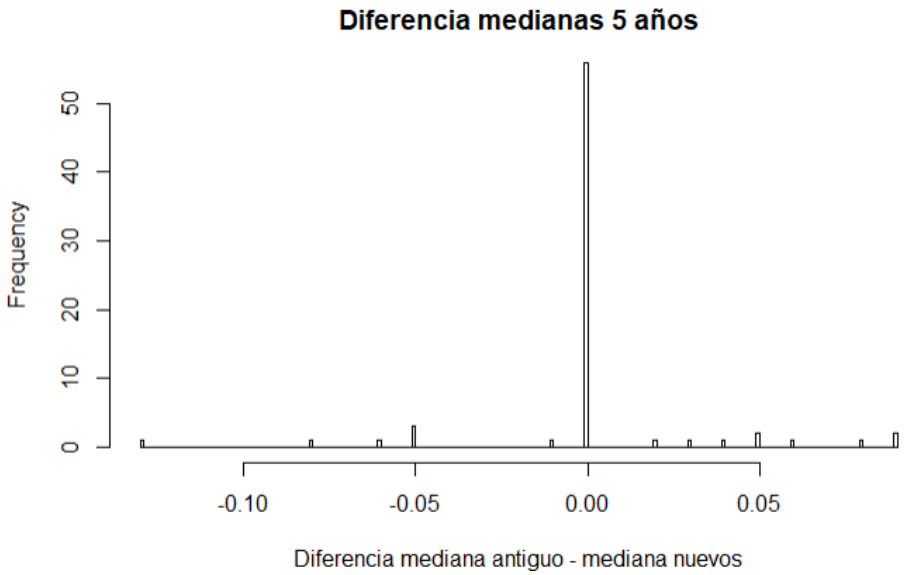

10.9
% Diferencia medianas consumos
-30
-25
-20
-15
-10
-5
0
0
20
40
60
80
diferencia edad (meses)

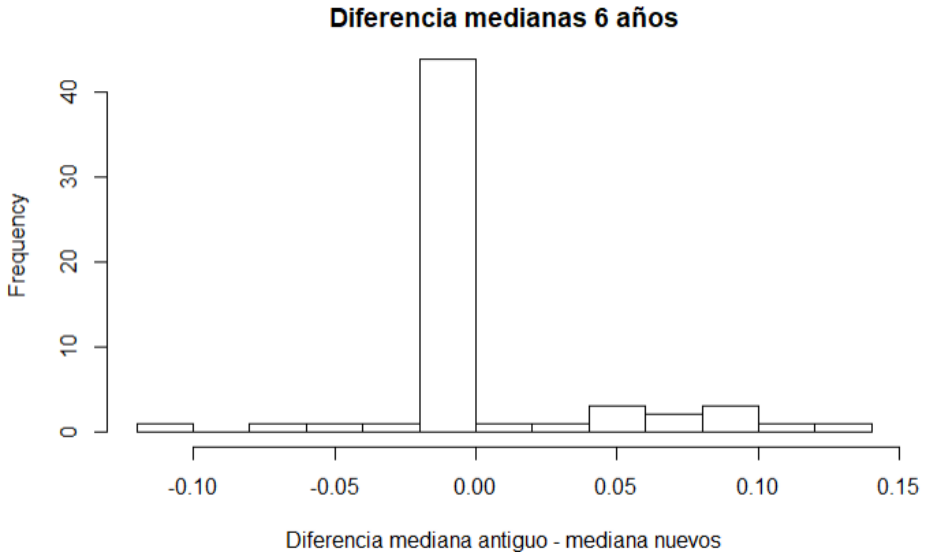

Diferencia medianas 1 año
Frequency
0
20
40
60
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
Diferencia mediana antiguo - mediana nuevos

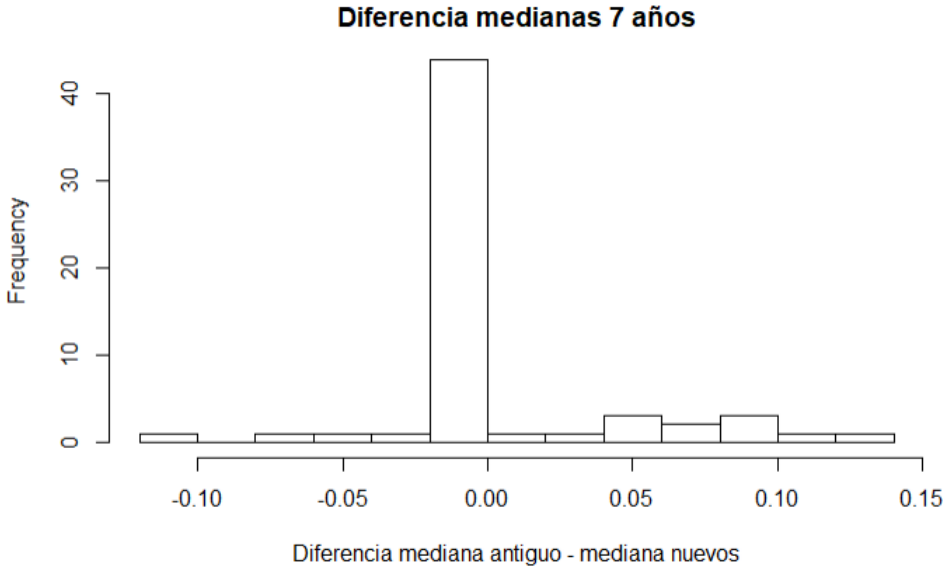

Diferencia medianas 2 años
Frequency
0
20
40
60
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
Diferencia mediana antiguo - mediana nuevos

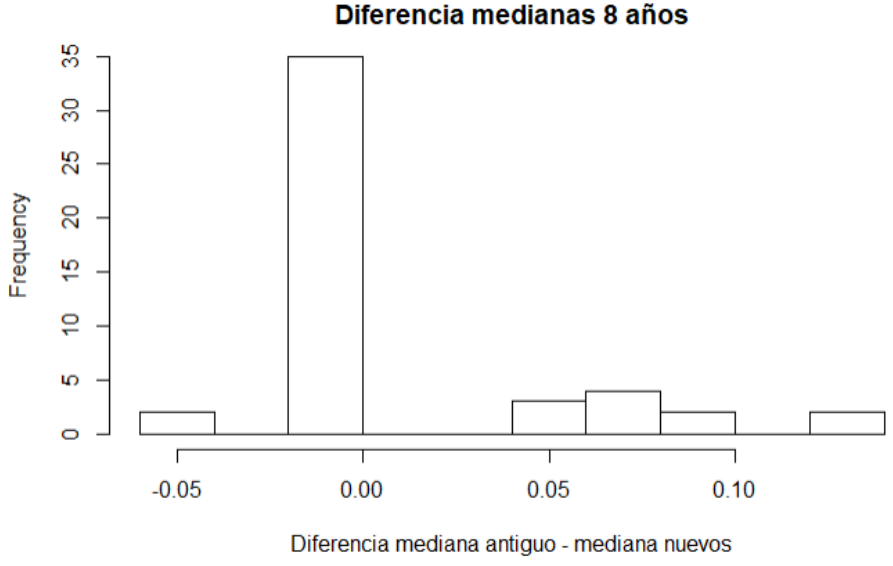

Diferencia medianas 3 año
Frequency
0
10
20
30
40
50
60
70
-0.15
-0.10
-0.05
0.00
0.05
0.10
Diferencia mediana antiguo - mediana nuevos


Diferencia medianas 4 años
Frequency
0
10
20
30
40
50
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
Diferencia mediana antiguo - mediana nuevos


Diferencia medianas 5 años
Frequency
0
10
20
30
40
50
-0.10
-0.05
0.00
0.05
Diferencia mediana antiguo - mediana nuevos

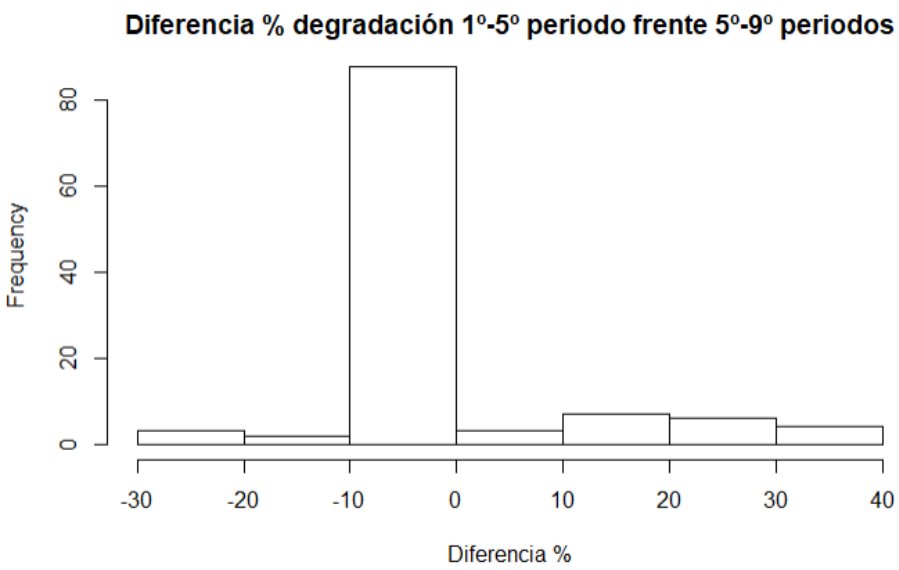

Diferencia medianas 6 años
Frequency
0
10
20
30
40
-0.10
-0.05
0.00
0.05
0.10
0.15
Diferencia mediana antiguo - mediana nuevos

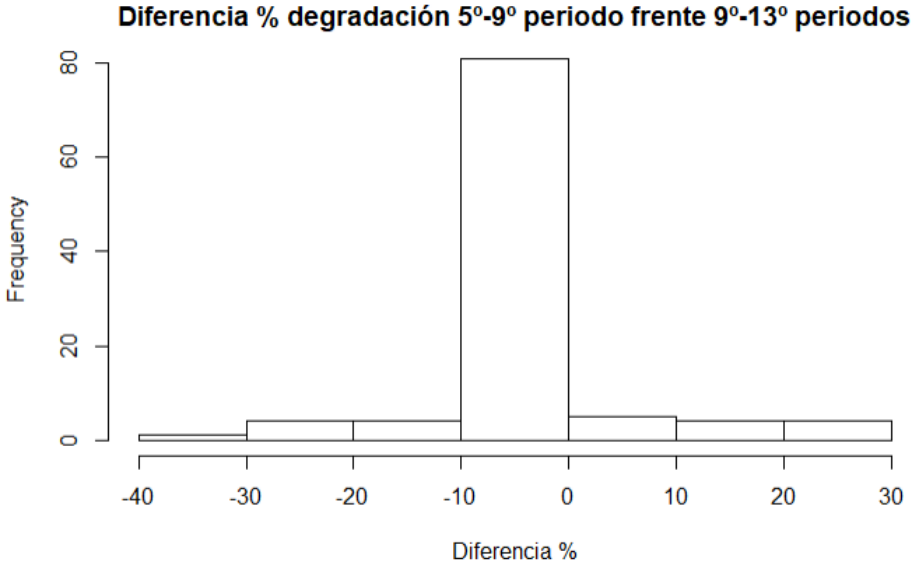

Diferencia medianas 7 años
Frequency
0
10
20
30
40
-0.10
-0.05
0.00
0.05
0.10
0.15
Diferencia mediana antiguo - mediana nuevos

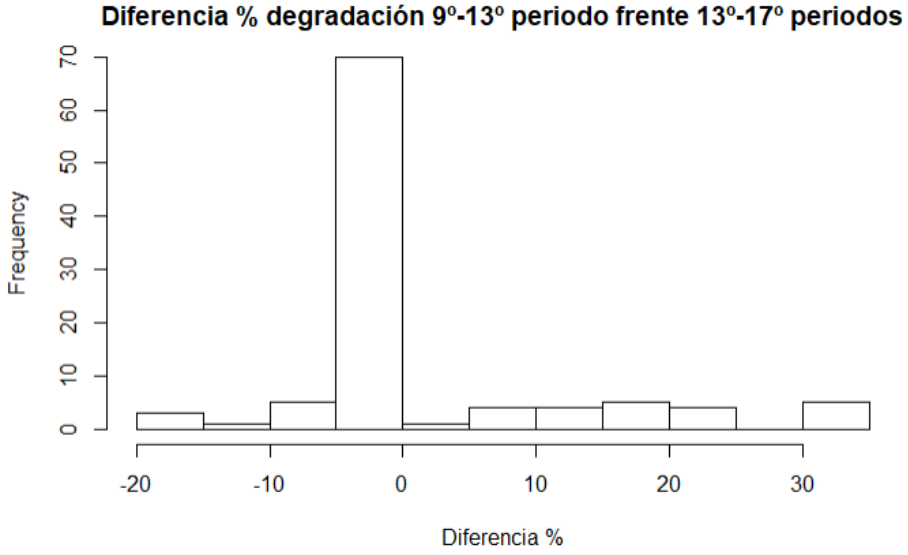

Diferencia medianas 8 años
Frequency
0
5
10
15
20
25
30
35
-0.05
0.00
0.05
0.10
Diferencia mediana antiguo - mediana nuevos

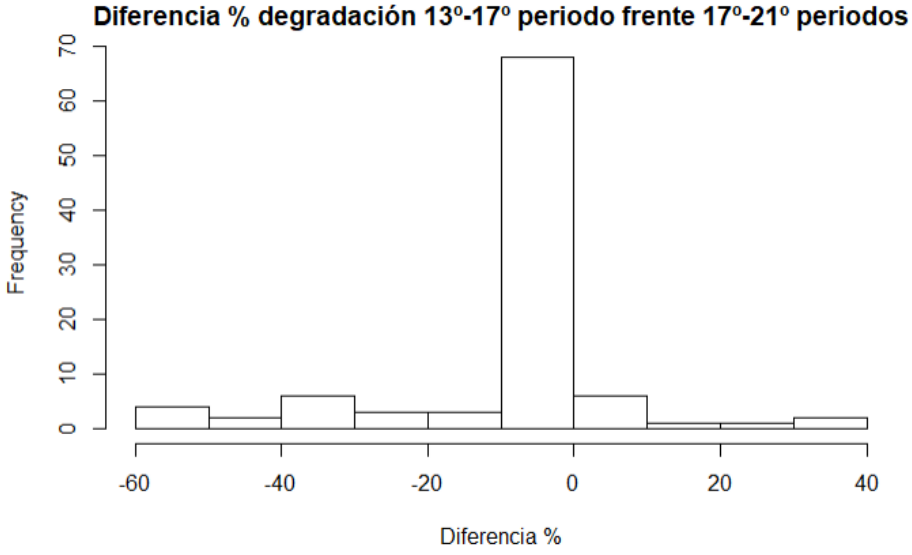
Anexo II

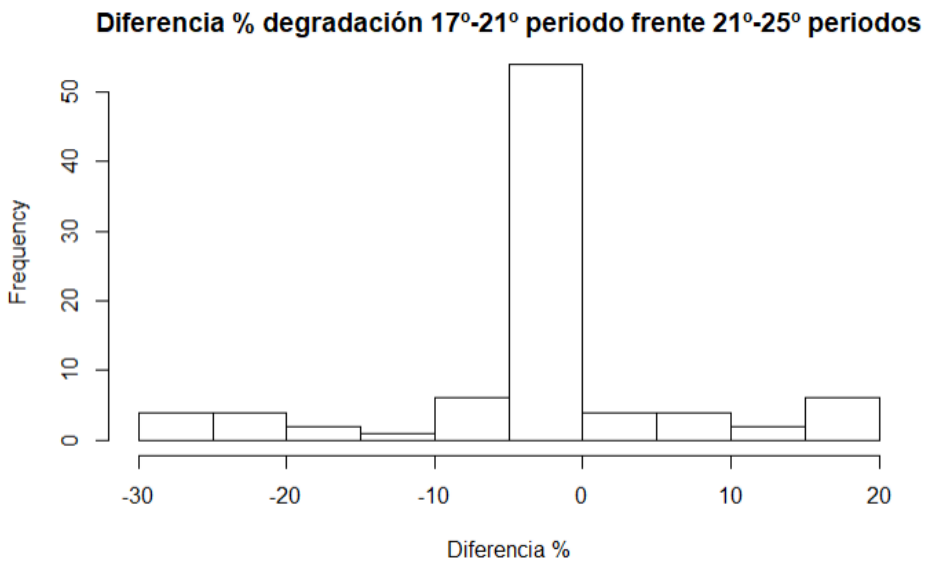
Segundo método de estudio.

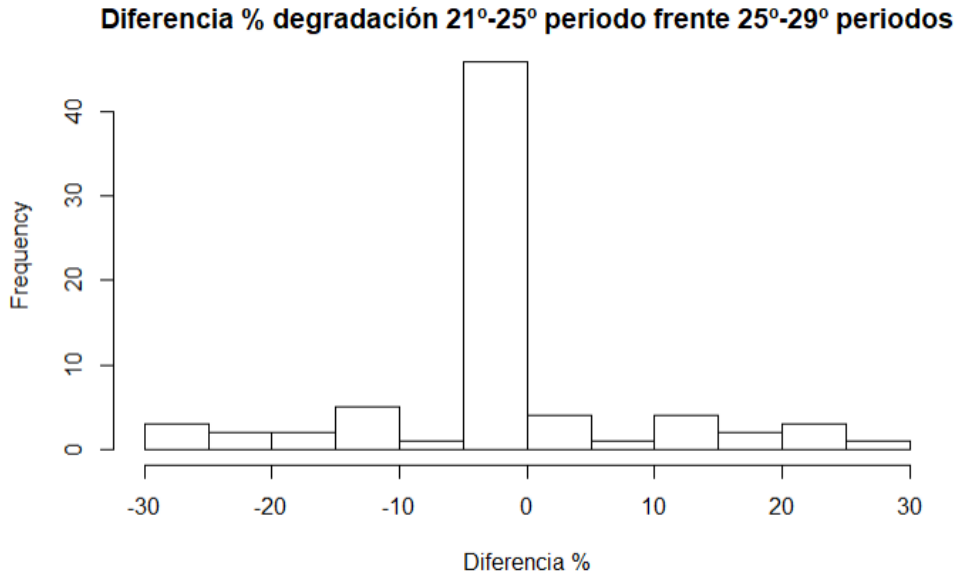

Diferencia % degradación 1º-5º periodo frente 5º-9º periodos
Frequency
0
20
40
60
80
-30
-20
-10
0
10
20
30
40
Diferencia %

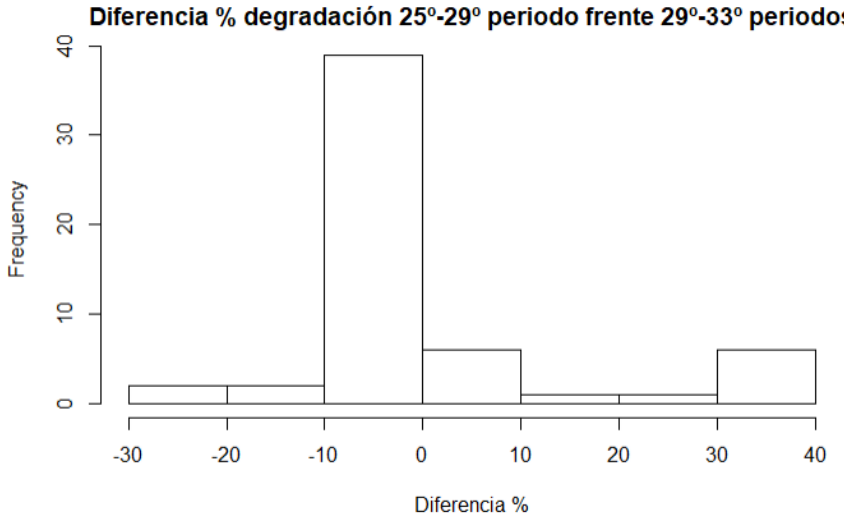

Diferencia % degradación 5º-9º periodo frente 9º-13º periodos
Frequency
0
20
40
60
80
-40
-30
-20
-10
0
10
20
30
Diferencia %


Diferencia % degradación 9º-13º periodo frente 13º-17º periodos
Frequency
0
10
20
30
40
50
60
70
-20
-10
0
10
20
30
Diferencia %


Diferencia % degradación 13º-17º periodo frente 17º-21º periodos
Frequency
0
10
20
30
40
50
60
70
-60
-40
-20
0
20
40
Diferencia %


Diferencia % degradación 17º-21º periodo frente 21º-25º periodos
Frequency
0
10
20
30
40
50
-30
-20
-10
0
10
20
Diferencia %


Diferencia % degradación 21º-25º periodo frente 25º-29º periodos
Frequency
0
10
20
30
40
-30
-20
-10
0
10
20
30
Diferencia %


Diferencia % degradación 25º-29º periodo frente 29º-33º periodos
Frequency
0
10
20
30
40
-30
-20
-10
0
10
20
30
40
Diferencia %

Anexo III

Programación en Rstudio

Librería para la lectura de datos en fichero Excel

```
library(readxl)

datos_2007 <- read_excel("C:/
```

Histogramas con discretización según *Scott,Freedman-Diaconis(FD) y Sturges*

```
par(mfrow=c(1,3)) # Nos sirve para la visualización de los 3 histogramas a la vez
hist(aux_hist,breaks = nclass.Sturges(aux_hist),main ="Sturges",xlab = "Consumo medio")
hist(aux_hist,breaks = nclass.scott(aux_hist),main = "Scott",xlab = "Consumo medio")
hist(aux_hist,breaks = nclass.FD(aux_hist),main = "FD",xlab = "Consumo medio")
```

Test de bondad de ajuste

```
install.packages("nortest") # Instalamos el paquete "nortest"
library(nortest) # Cargamos la librería "nortest"

quantile(aux[4])-quantile(aux[2])/sd(aux)
lillie.test(aux) # test de K-s con correción Lillie
ad.test(aux) # test anderson-Darling
cvm.test(aux) # Cramer-von Mises
sf.test(aux) # saphiro-Francia test
```

Cálculo medianas primeros periodos

Realizamos una función principal a la que le pasaremos un listado con los datos de consumo y las fechas de instalación. Su salida nos ofrece la cantidad de datos antes y después de eliminar los que están fuera de intervalo y sus respectivas medianas. El nombre de esta función es *función_medianas_periodo1*.

```
funcion_mediana_periodo1=function(listado,nombres){


  cantidad_original=NULL;mediana_original=NULL;cantidad_filtrada=NULL;mediana_filtrada=NULL;
salida=NULL;
  for (i in 1:length(listado)) { # La longitud del listado será los años donde tengamos
datos, en este caso 10 años

    for (j in 1:12) { # 12 meses

      ## Tomamos los valores del primer periodo del año i y mes j y los llamamos aux, con
algunas condiciones
      aux=listado[[i]]$`Periodo 1`[listado[[i]]$Mes==j & listado[[i]]$`Periodo 1`>0 &
!is.na(listado[[i]]$`Periodo 1`)]
```

```
 # Calculamos mediana original y cantidad de datos original
 cantidad_original=length(aux)  ## floor para quitar decimales
 #mediana_original=round(mean(aux,na.rm = TRUE),3) ## round para poner 3 decimales
 mediana_original=median(aux,na.rm = TRUE)
```

```
      # Filtramos los datos por la regla empírica tantas veces como haga falta

       while (length(which(aux<(mean(aux,na.rm = T)-3*sd(aux,na.rm = T)) |
aux>(mean(aux,na.rm = T)+3*sd(aux,na.rm = T))))>0)
         {
       aux=aux[-c(which(aux<(mean(aux,na.rm = T)-3*sd(aux,na.rm = T)) | aux>(mean(aux,na.rm
= T)+3*sd(aux,na.rm = T))))]
       }
```

```
      # Calculamos mediana filtrada y cantidad de datos ya filtrados, añadimos modificacion
if, si hay menos de 30 datos la mediana la   bajamos a 0

      cantidad_filtrada=length(aux)
      #media_filtrada=round(mean(aux,na.rm=T),3) # otra manera de hacerlo
      mediana_filtrada=median(aux,na.rm=T) # otra manera de hacerlo

      if(cantidad_filtrada>30){
        mediana_filtrada=round(median(aux,na.rm=T),3)
      }else{mediana_filtrada=0}
```

```
      # Preparamos los datos para la salida
      cantidad_original=floor(cantidad_original)
      cantidad_filtrada=floor(cantidad_filtrada)
      media_original=round(mediana_original,3)
      media_filtrada=round(mediana_filtrada,3)

      aux1=c(cantidad_original,mediana_original,cantidad_filtrada,mediana_filtrada)
      salida=cbind(salida,aux1)
    }
  }

  ## Pongo nombres a columnas y filas
    colnames(salida)=c(nombres)
    rownames(salida)=c("cantidad original","mediana original","cantidad filtrada","mediana
filtrada")

  return(salida)

}
```

Para llamar a la anterior función tomamos

```
# Listado con los datos de consumo por año

listado=list(datos_2007,datos_2008,datos_2009,datos_2010,datos_2011,datos_2012,datos_2013,da
tos_2014,datos_2015,datos_2016,datos_2017);

# Nombres para el listado de salida

nombres=c("1.7","2.7","3.7","4.7","5.7","6.7","7.7","8.7","9.7","10.7","11.7","12.7","1.8","
2.8","3.8","4.8","5.8","6.8","7.8","8.8","9.8","10.8","11.8","12.8","1.9","2.9","3.9","4.9",
"5.9","6.9","7.9","8.9","9.9","10.9","11.9","12.9","1.10","2.10","3.10","4.10","5.10","6.10"
,"7.10","8.10","9.10","10.10","11.10","12.10","1.11","2.11","3.11","4.11","5.11","6.11","7.1
1","8.11","9.11","10.11","11.11","12.11","1.12","2.12","3.12","4.12","5.12","6.12","7.12","8
.12","9.12","10.12","11.12","12.12","1.13","2.13","3.13","4.13","5.13","6.13","7.13","8.13",
"9.13","10.13","11.13","12.13","1.14","2.14","3.14","4.14","5.14","6.14","7.14","8.14","9.14
","10.14","11.14","12.14","1.15","2.15","3.15","4.15","5.15","6.15","7.15","8.15","9.15","10
.15","11.15","12.15","1.16","2.16","3.16","4.16","5.16","6.16","7.16","8.16","9.16","10.16",
"11.16","12.16","1.17","2.17","3.17","4.17","5.17","6.17","7.17","8.17","9.17","10.17","11.1
7","12.17")

# Dibujamos los valores obtenidos

plot(funcion_mediana_periodo1(listado,nombres)[4,],type = "b",xlab = " Tiempo (meses)
Correspondencia contador instalado",ylab = "Mediana del consumo de primeros periodos")
```

Estimar consumos nulos

Partiendo de la tabla de primeros consumos, guardamos en un vector las medias con los datos ya filtrados.

```
# Vector con la mediana de primeros consumos

primeros_consumos=funcion_mediana_periodo1(listado,nombres)[4,]

# Gráfica vector media de primeros consumos

plot(primeros_consumos,type = "l")
```

```
for (i in 1:length(primeros_consumos)) { # Sustituimos los valores nulos

  if(primeros_consumos[[i]]==0 &
i<12){primeros_consumos[[i]]=(primeros_consumos[[i-1]]+primeros_consumos[[i+1]])/2}
  if(primeros_consumos[[i]]==0 & i>12){primeros_consumos[[i]]=primeros_consumos[[i-12]]}
}
```

Dibujamos ambas gráficas conjuntas con plot y points

```
plot(funcion_mediana_periodo1(listado,nombres)[4,],type="l");points(primeros_consumos,type =
"l",col="green")
```

Cálculo medias para periodos sucesivos correspondientes a muestras de contadores por mes de instalación

Igualmente haremos uso de una función, en este caso llamada *función_medianas_periodo_serie*

```
funcion_medianas_periodo_serie=function(listado,nombres){

  aux=NULL;aux1=NULL;salida=NULL;salida=as.list(salida); # auxiliares de cálculo

  for (i in 1:length(listado)) { # dependiente del tamaño del listado

   salida_aux=NULL;salida_aux=as.list(salida_aux);
```

```
   for (j in 1:12) {

     mediana_filtrada=NULL;cantidad_datos=NULL;

     ## creamos dataframe con datos del 2007, 2008...clasificandolos por meses, el primero
2007 mes 1
     aux=data.frame(listado[[i]][listado[[i]]$Mes==j & listado[[i]]$`Periodo 1`>0 &
listado[[i]]$`Periodo 1`<10,])

     ## Hallamos los periodos disponibles para esos datos contando las filas "numericas"
     clase=-7;for (q in 1:length(aux)) {if(class(aux[[q]])=="numeric"){clase=clase+1}}
```

```
    ## Lo que queda es calcular la media de todos los periodos

    for (h in 1:clase) {

      aux1=aux[[8+h]] ## Ya que los periodos entran a partir de la columna 9

      ## Filtramos los datos por regla empírica

      while (length(which(aux1<(mean(aux1,na.rm = T)-3*sd(aux1,na.rm = T)) |
aux1>(mean(aux1,na.rm = T)+3*sd(aux1,na.rm = T))))>0)
        {
         aux1=aux1[-c(which(aux1<(mean(aux1,na.rm = T)-3*sd(aux1,na.rm = T)) |
aux1>(mean(aux1,na.rm = T)+3*sd(aux1,na.rm = T))))]
       }
```

```
      ## creamos vectores con las medianas de los diferentes periodos,añado modificacion si
cantidad datos < 30, media 0

      cantidad_datos=c(cantidad_datos,length(aux1))

      if(length(aux1)>30){

        mediana_filtrada=c(mediana_filtrada,median(aux1,na.rm = T)[!is.na(median(aux1,na.rm
= T))])

        }else{mediana_filtrada=c(mediana_filtrada,0)}

    }
```

```
    # Preparamos salida auxiliar

      salida_aux[[j]]=rbind(mediana_filtrada)
  }

   # Esta salida auxiliar la vamos almacenando para la salida general
   names(salida_aux)=c(nombres[1:length(salida_aux)])

   #nombres para facilitar lectura de la tabla
   nombres=nombres[13:length(nombres)]
   salida[[i]]=salida_aux

 }
 salida

 return(salida)


}
```

Medias móviles

Presentamos la obtención de medias móviles. Se ha realizado media móvil centrada, con lo cual, en las gráficas, al principio y al final perderíamos datos. Para evitarlo se realiza un sumatorio de los datos de principio y final para mejorar la visualización.

```
# Estacionalidad de 12 medidas

funcion_media_movil_mensual=function(serie_temporal){
  salida_mensual=NULL;
  for (i in 1:length(serie_temporal)) {

    if(i<=12){salida_mensual[i]=(sum(serie_temporal[1:i]))/i}

    if(i>12){salida_mensual[i]=(sum(serie_temporal[(i-11):i]))/12}
  }

  return(salida_mensual)
}
```

```
# Estacionalidad de 4 medidas

funcion_media_movil_trimestral=function(serie_temporal){
  salida_trimestral=NULL;
  for (j in 1:length(serie_temporal)) {

    if(j<=4){salida_trimestral[j]=(sum(serie_temporal[1:j]))/j}

    if(j>4){salida_trimestral[j]=(sum(serie_temporal[(j-3):j]))/4}
  }

  return(salida_trimestral)
}
```

Método K-NN

```
# Necesitaremos librería class
library("class", lib.loc="C:/Program Files/R/R-3.5.0/library")
```

Leemos datos y eliminamos todas aquellas viviendas que no contengan datos en todos los periodos de estudio

```
# leemos datos 07
enero07=data.frame(datos_2007[datos_2007$Mes==1 & datos_2007$`Periodo 1`>0,])
#Eliminamos los periodos fuera de los 6 años de estudio
enero07=data.frame(enero07[,21:40])
# Eliminamos filas que no estén completas
for (i in 1:20) {

  if(length(which(is.na(enero07[,i])))>0){
  enero07=enero07[-c(which(is.na(enero07[,i]))),]
  }
}
```

Para realizar muestras aleatorias en las viviendas se ha utilizado el comando “sample”

```
# Enero 2007
# Realizamos muestra aleatoria para elección viviendas de prueba
aux = sample(x=c(1:length(enero07$Periodo.13)),
             replace=FALSE, size=35)

# Guardamos esas viviendas de entrenamiento
entrenamiento07=enero07[c(aux),]
```

```
# Guardamos data.frame con las viviendas restantes
enero07r=enero07[-c(aux),]

# Realizamos muestra aleatoria para elección viviendas de estudio
aux=sample(x=c(1:length(enero07r$Periodo.13)),
           replace=FALSE, size=35)

#Guardamos esas viviendas de estudio
estudio07=enero07r[c(aux),]
```

Adecuamos los datos para insertarlos en la función knn

```
# Adecuamos los datos para introducirlos en la funcion Knn, primero los de entrenamiento

nombres_entreno=rep("periodo",time=20)
names(entrenamiento07)=nombres_entreno
names(entrenamiento08)=nombres_entreno
names(entrenamiento09)=nombres_entreno
names(entrenamiento10)=nombres_entreno

entrenamiento=rbind(entrenamiento07,entrenamiento08,entrenamiento09,entrenamiento10)
```

```
# Guardamos a qué año pertenece cada uno

nombre_cl=c(rep("2007",35),rep("2008",35),rep("2009",35),rep("2010",35))

# Pasamos ahora a los datos de prueba

names(estudio07)=nombres_entreno
names(estudio08)=nombres_entreno
names(estudio09)=nombres_entreno
names(estudio10)=nombres_entreno

estudio=rbind(estudio07,estudio08,estudio09,estudio10)
```

Insertamos los datos en función knn y obtenemos los resultados

```
# Realizaremos método Knn para diferentes k y veremos resultados

knn(entrenamiento, estudio, nombre_cl, k = 20, prob=TRUE)
aux1=knn(entrenamiento, estudio, nombre_cl, k = 20, prob=FALSE)


# Vemos cuanos ha acertado

which(aux1[1:35]=="2007")
which(aux1[36:70]=="2008")
which(aux1[71:105]=="2009")
which(aux1[106:140]=="2010")
```

Contraste mediante prueba U de Mann-Whitney

Este contraste se realiza por la función *funcionavsn*, a la cual se envía los datos para ser comparados, y ella se encarga de eliminar valores atípicos, hacer la comparación de las muestras mediante la realización de la prueba U de Mann-Whitney.

Analiza esta prueba y determina si hay diferencia entre las muestras y la cuantifica, dando un intervalo de confianza para la diferencia.

```
funcionAvsN=function(dato_antiguo,dato_nuevo){


  ## Filtrado de datos

 while (length(which(dato_antiguo<(mean(dato_antiguo)-3*sd(dato_antiguo)) |
dato_antiguo>(mean(dato_antiguo)+3*sd(dato_antiguo))))>0)
 {
 dato_antiguo=dato_antiguo[-c(which(dato_antiguo<(mean(dato_antiguo)-3*sd(dato_antiguo)) |
dato_antiguo>(mean(dato_antiguo)+3*sd(dato_antiguo))))]
 }
```

```
 while (length(which(dato_nuevo<(mean(dato_nuevo)-3*sd(dato_nuevo)) |
dato_nuevo>(mean(dato_nuevo)+3*sd(dato_nuevo))))>0)
 {
 dato_nuevo=dato_nuevo[-c(which(dato_nuevo<(mean(dato_nuevo)-3*sd(dato_nuevo)) |
dato_nuevo>(mean(dato_nuevo)+3*sd(dato_nuevo))))]

 }
  ## Analizamos p-valor y hallamos diferencias de medianas

modelo=c(dato_antiguo,dato_nuevo)
tiempo=c((rep(x="antiguo", times=length(dato_antiguo))),(rep(x="nuevo", times=length(dato_nuevo))))
   ## Analizamos p-valor y hallamos diferencias de medianas

modelo=c(dato_antiguo,dato_nuevo)
tiempo=c((rep(x="antiguo", times=length(dato_antiguo))),(rep(x="nuevo", times=length(dato_nuevo))))
if(length(dato_nuevo)>=30 & length(dato_antiguo)>=30){

      aux_wilcox=wilcox.test(modelo~tiempo)[3]
      aux_wilcox=as.numeric(aux_wilcox)

  if(wilcox.test(modelo~tiempo)[3]<0.05){


  zm=median(dato_antiguo)-median(dato_nuevo)

  e_medio=zm/median(dato_nuevo)

  if( zm<=0 ) {
 salida_funcion=c(length(dato_antiguo),length(dato_nuevo),round(aux_wilcox,7),round(zm,2),round(e_medio*
100,2))

 }else
   {

   zm=0;e_medio=0
   salida_funcion=c(length(dato_antiguo),length(dato_nuevo),round(aux_wilcox,7),round(zm,2),round(e_medi
o*100,2))
  }
return(salida_funcion)

}
```

Implementación primer método

Al tener muchas hojas de datos con muchos periodos se implementó una función llamada *funcion_Mann_Whitney* para comparar estas bases de datos. Esta función hacía llamadas a la función anterior *funcionavsn* y guardaba los resultados de todas las comparaciones de una base de datos entre los periodos nuevos y los sucesivos. Vamos a verla por pasos:

Primero cargamos todos los datos en nuestro programa provenientes de hojas Excel

```
library(readxl)

datos_2007 <- read_excel("C:/
datos_2007$Dia=as.numeric(datos_2007$Dia)
datos_2007$Mes=as.numeric(datos_2007$Mes)

datos_2008 <- read_excel("C:/
datos_2008$Dia=as.numeric(datos_2008$Dia)
datos_2008$Mes=as.numeric(datos_2008$Mes)
```

Se han eliminado ciertos nombres por privacidad de los datos, aquí hemos cargado los datos correspondientes a los años 2007 y 2008, esto se realizó con datos hasta 2017 que eran los disponibles.

Tomamos nuevas variables donde guardamos los datos por mes

```
m=c(1,4,7,10) ## Mes para la comparación, siguientes:(2,5,8,11) ( 3,6,9,12)
d=c(0,31) ## Días pra la comparación


datos1.7=data.frame(datos_2007[datos_2007$Mes==m[1] & datos_2007$Dia>d[1] & datos_2007$Dia<=d[2]  &
datos_2007$`Periodo 1`>0,])
datos1.7=(datos1.7[,9:53])

datos4.7=data.frame(datos_2007[datos_2007$Mes==m[2] & datos_2007$Dia>d[1] & datos_2007$Dia<=d[2]  &
datos_2007$`Periodo 1`>0,])
datos4.7=(datos4.7[,9:53])
```

Hemos dejado la opción de poder incluso seleccionar por días mediante la variable “d”.

Los datos anteriores realmente son tablas de periodos sucesivos para un mismo contador.

Lo siguiente que se realiza es un listado de listados con todas estas hojas

```
datos_nuevos=list(datos1.7,datos4.7,datos7.7,datos10.4,datos1.8,datos4.8,datos7.8,dato
```

Y después guardamos en otra variable un listado de los primeros consumos

```
datos_aux=list(datos_2007$`Periodo 1`[datos_2007$Mes==m[2] & datos_2007$Dia>d[1] & datos_2007$Dia<=d[2]
& datos_2007$`Periodo 1`>0],datos_2007$`Periodo 1`[datos_2007$Mes==m[3] & datos_2007$Dia>d[1] &
datos_2007$Dia<=d[2]  & datos_2007$`Periodo 1`>0],datos_2007$`Periodo 1`[datos_2007$Mes==m[4] &
```

También tomamos dos variables de tiempo y nombres para la salida de datos

```
nombres=c("1.7","4.7","7.7","10.7","1.8","4.8","7.8","10.8","1.9","4.9","7.9","10.9","1.10","4.10","7.10"
,"10.10","1.11","4.11","7.11","10.11","1.12","4.12","7.12","10.12","1.13","4.13","7.13","10.13","1.14","4
.14","7.14","10.14","1.15","4.15","7.15","10.15","1.16","4.16","7.16","10.16","1.17","4.17","7.17","10.17
")
```

Con todo esto, creamos una función principal para ir comparando todos los periodos de un contador con periodos de contadores recién instalados

```
funcion_principal=function(listado1,listado2,nombres_comparacion){

  q=NULL;datos_comparacion=NULL;aux=NULL;aux2=NULL;
```

```
 for (q in 1:length(listado2)) {
   a=NULL;b=NULL;
   a=listado1[,q+1][!is.na(listado1[,q+1]) & listado1[,q+1]>0]
   b=listado2[[q]][!is.na(listado2[[q]]) & listado2[[q]]>0]
   aux=c(funcionAvsN(a,b))
   aux2=(c(aux[[1]],aux[[2]],aux[[3]],aux[[4]],aux[[5]]))
   datos_comparacion=rbind(datos_comparacion,aux2)

 }
```

```
  row.names(datos_comparacion)=nombres_comparacion
  colnames(datos_comparacion)=c("dato_antiguo","dato_nuevo","p-valor","Za - Zn","e_medio");
   return(datos_comparacion)

}
```

Como complemento a la anterior función, también se ha creado una función de dibujo para ir graficando los valores de la tabla de salida

```
funcion_dibujo=function(tabla_valores,nombre_dato,tiempo){

  j=NULL;aux_grafico=NULL;error=NULL;
  for (j in 1:length(tabla_valores[,1])) {

      aux_grafico=(tabla_valores[j,4])
      aux_grafico=as.numeric(aux_grafico)
      aux_grafico=aux_grafico
      error=c(error,aux_grafico)
}

  return(plot(x=tiempo,y=error,xlab = "diferencia edad (meses)",ylab = "% Diferencia medianas
consumos",main = nombre_dato,type = "b"))
}
```

Implementación del segundo método

Este segundo método se ha implementado de manera similar al primero, únicamente tomando las muestras a comparar del mismo mes de instalación y diferenciando periodos

Bibliografía y fuentes consultadas

- *Análisis de series temporales, David peña [4]*

- *Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). The New S Language. Wadsworth & Brooks/Cole. [11]*

- *Birgé, L.; Rozenholc, Y. (2006). "How many bins should be put in a regular histogram" [12]*

- *Empresa FACSA https://www.facsa.com/la-empresa/ [3]*

- *Estudio del comportamiento metrológico de los contadores en abastecimientos de agua. Optimización de su gestión para la reducción de las pérdidas comerciales. F. Gavara (2015), [2]*

- *Freedman, D. and Diaconis, P. (1981) On this histogram as a density estimator: L2 theory. Zeit. Wahr. ver. Geb. [15]*

- *Probabilidad y estadística para ingeniería y ciencia, William Mendenhall & Terry Sincich[6]*

- *Rstudio https://www.rstudio.com[16]*

- *Scott, D.W. (1979) On optimal and data-based histograms. Biometrika[13]*

- *Sturges, H. (1926) The choice of a class-interval. J. Amer. Statist. Assoc., [14]*

- *Universidad de Valencia https://www.uv.es/webgid/Descriptiva/tema_5_correlacin.html[10]*

- *Universidad del País Vasco http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t9knn.pdf[8]*

- *Universidad Carlos III de Madrid http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/KNNyPrototipos.pdf[9]*

- *Universidad de Barcelona http://www.ub.edu/aplica_infor/spss/cap6-2.htm[7]*

- *Universidad de granada http://wpd.ugr.es/~bioestad/guia-r-studio/practica-1-r-studio/[1]*

- *Universidad de valencia https://www.uv.es/webgid/Inferencial.html[5]*