En este espacio vamos a redactar todo lo relacionado con la limpieza de datos atípicos, es importante conocer que se han tratado de forma diferente según el mes.

**La norma para considerar un dato atípico ha sido:**

Generalmente, valores mayores a 1000 KW no pueden darse en 15 minutos por lo tanto son atípicos

No obstante esta columna la hemos creado nosotros, cada valor depende del que hubiera en la fila anterior, y dado que los sensores que calculan el KWH pueden fallar, hay ocasiones donde no registraban los datos durante un tiempo, que puede oscilar entre sólo 15 minutos hasta incluso un día entero (un ejemplo de esto es el 8 de mayo).

Aunque no registran los datos, se van acumulando durante ese tiempo y se suman y calculan en el próximo dato que registra por lo que hay ocasiones donde el valor de KWH/15 min no es de 15 minutos sino que es de más tiempo.

Si durante 15 minutos registra de media 10 KW, si deja de registrar 5 horas, en el próximo dato habrá 300 KW, (puede que no se registren todos los datos, de normal habrán pérdidas porque también falla al calcular, no obstante es una buena aproximación) en algunos edificios este consumo cada 15 min puede llegar a ser hasta de 200 o 300 *(Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M) Por lo que datos que superen los 10000 - 20000 KWH/15min son posibles, pero también técnicamente verdaderos*

No hemos considerado atípicos estos datos, esos datos aunque no reflejan el consumo durante 15 minutos, si lo hacen para todo el periodo acumulado que es lo que buscábamos al crear la columna, y para el estudio que queremos hacer nosotros, los hemos dado por válidos.

Los valores que no hemos dado por válidos son otros que excedían absurdamente, como un consumo de un 1.600.000 KWH en 15 minutos, se puede comprobar que en este caso no es por fallo de los sensores anteriores en el excel proporcionado por el departamento de medio ambiente.

**PARTE 1 (SEPTIEMBRE - FEBRERO) 2018-2019**

**En los siguientes 6 meses hemos procedido de manera diferente a la eliminación de datos atípicos que en el resto puesto que en estos encontramos una gran cantidad. En este caso hemos seguido los siguientes pasos:**

1º Se ha identificado los datos atípicos de KWH/15 MIN mediante un gráfico de aberrantes en el programa Statgraphics.

2º Se crea una tabla que demuestra la gran diferencia entre la media y la mediana y por consiguiente afirma la presencia de datos muy grandes.

3º Se hace la operación media más dos desviaciones típicas.

4º Se eliminan los datos más grandes del resultado de la operación mediante Python.

5º Se repite el proceso hasta el punto 2. Si la media sigue siendo muy diferente a la mediana pasamos a los puntos 3 y 4.

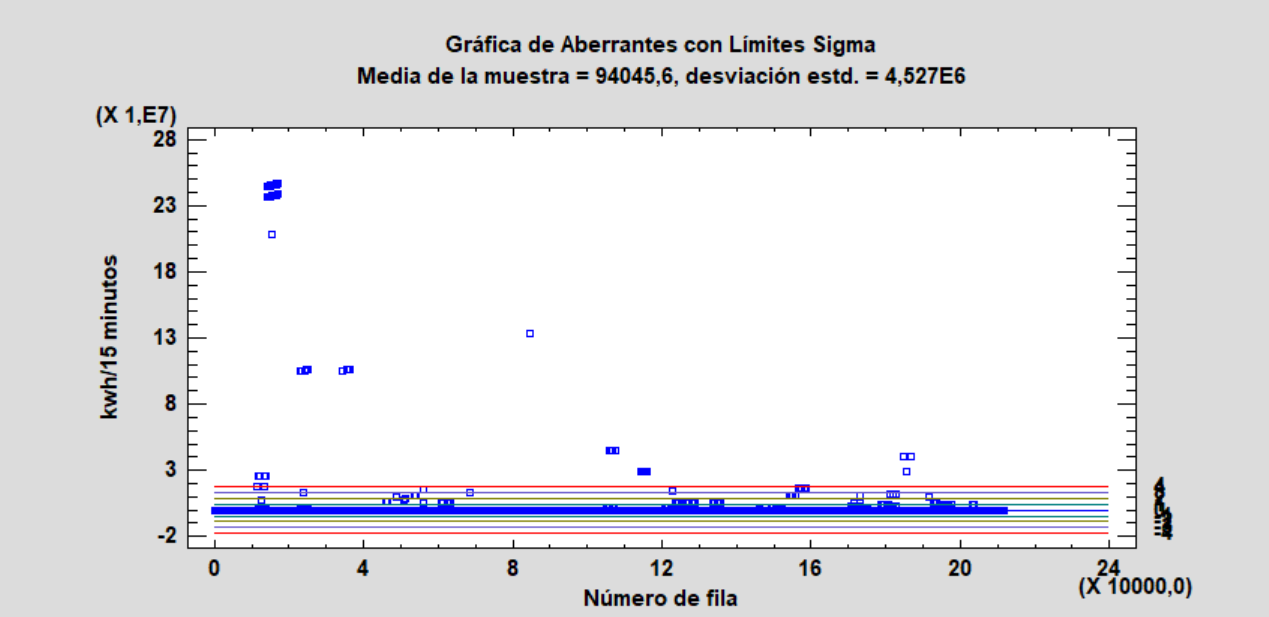
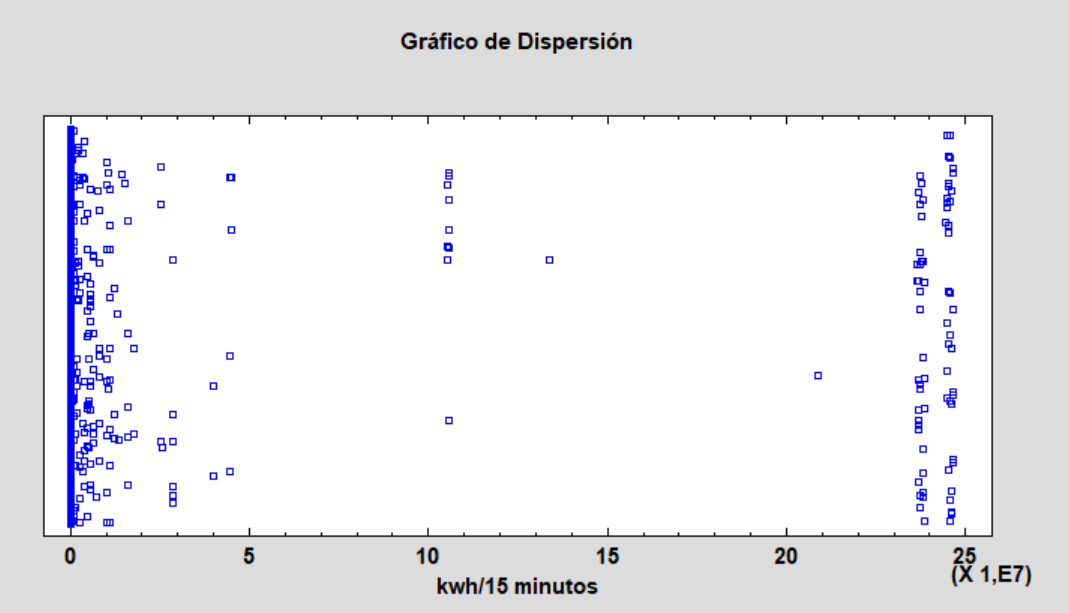
La columna creada KWH/15 MIN en ocasiones muestra el formato .0, por ejemplo, 4.0, 6.0, 27.0, etc. En estos meses Statgraphics multiplica por 10 el número dejándolo como 40, 60, 27 … Luego, en estos casos seguiremos haciendo la tabla pero dividiremos los número entre 10, puesto que si no lo hacemos estaríamos introduciendo una cantidad errónea en Python, que sí que interpreta el 4.0,6.0,27.0 como un 4,6 y 27.

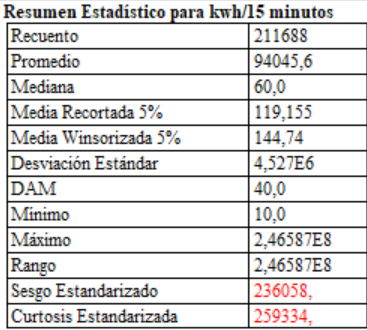
A continuación, mostraremos la evolución de estos meses mediante gráficos y tablas antes y después de limpiar estos datos atípicos.

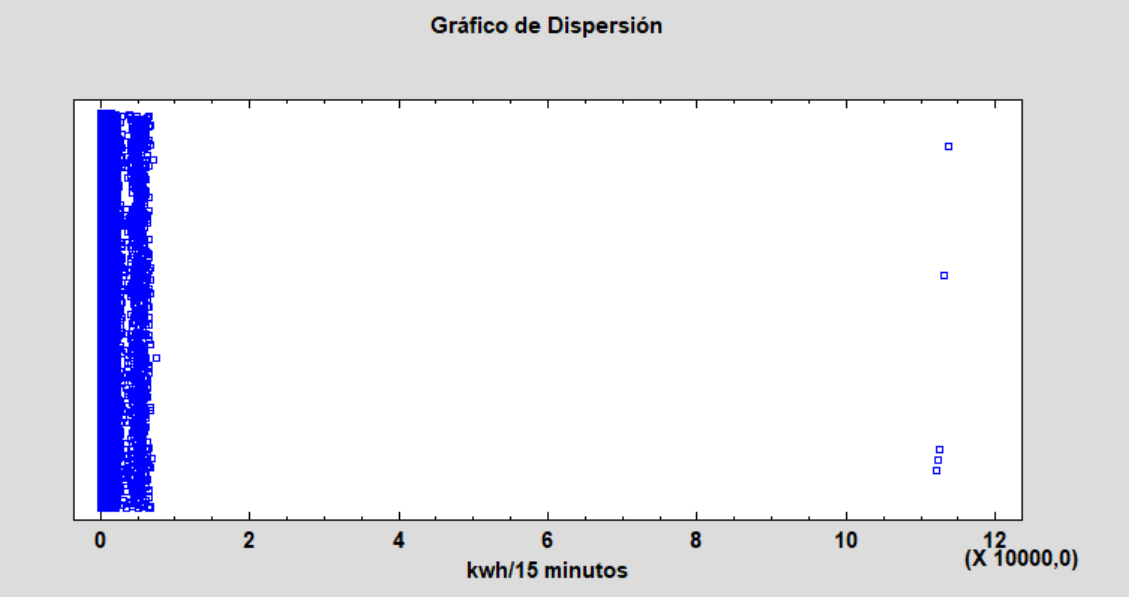
Dejaremos el código en el anexo, con información detallada de lo que hemos hecho en Python.

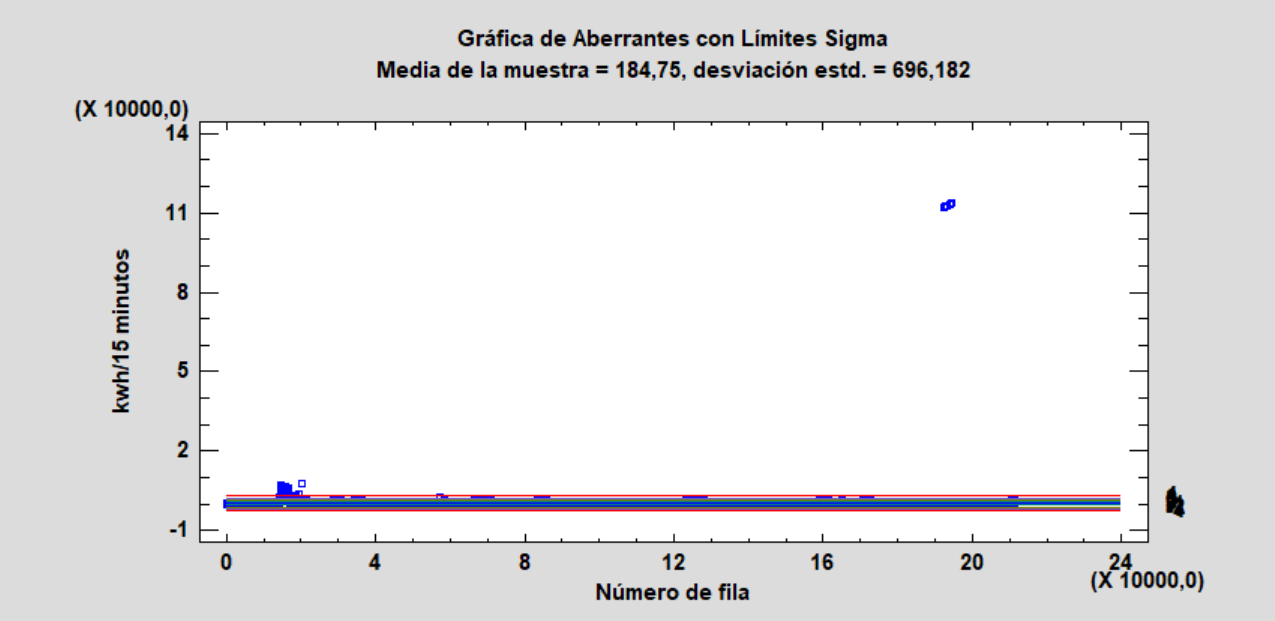
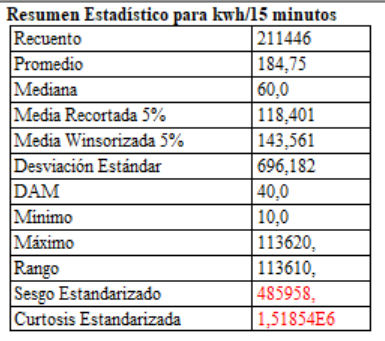
**Septiembre**

**Antes de limpiar**

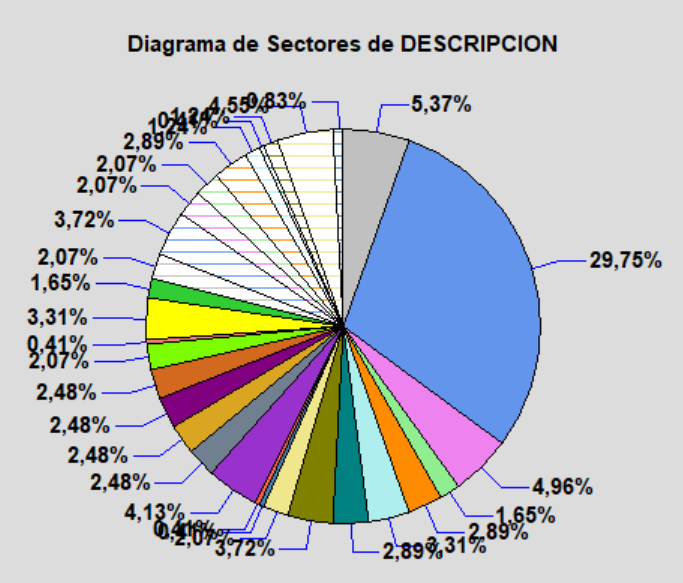
****

****

**Después de limpiar**

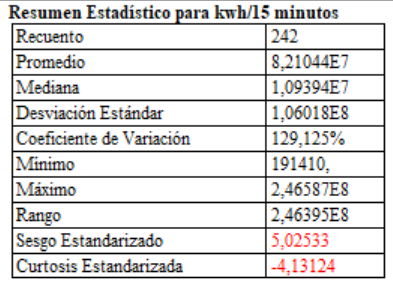
****

**Sep Atípicos ‘ Analizamos sólo los atípicos’ Para ver en qué edificio se produce este error con más frecuencia.**

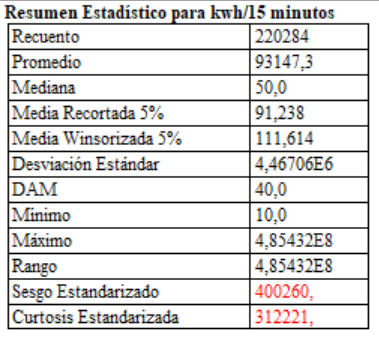
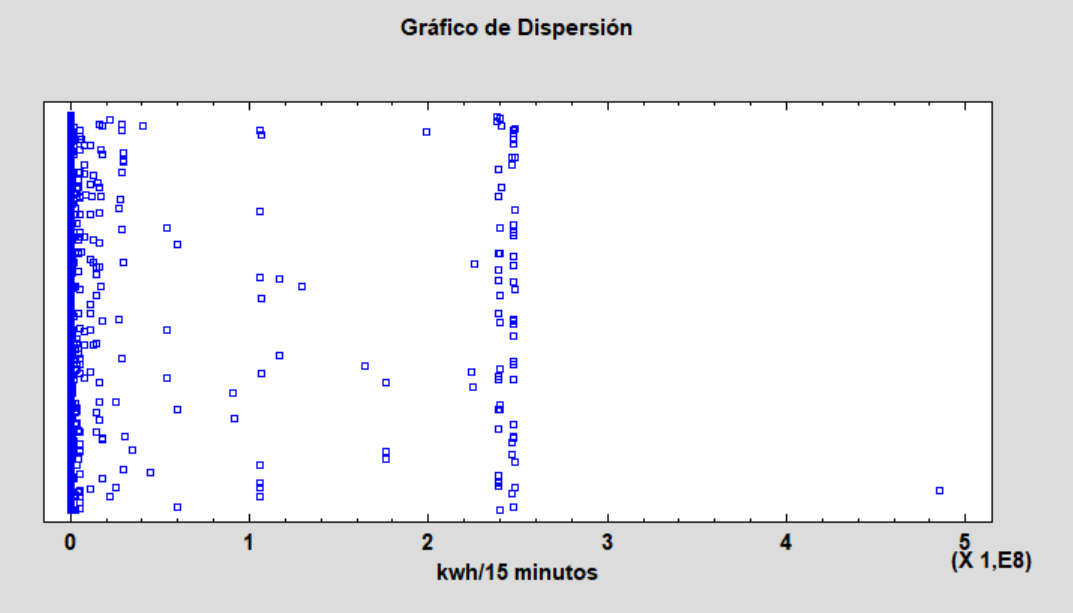
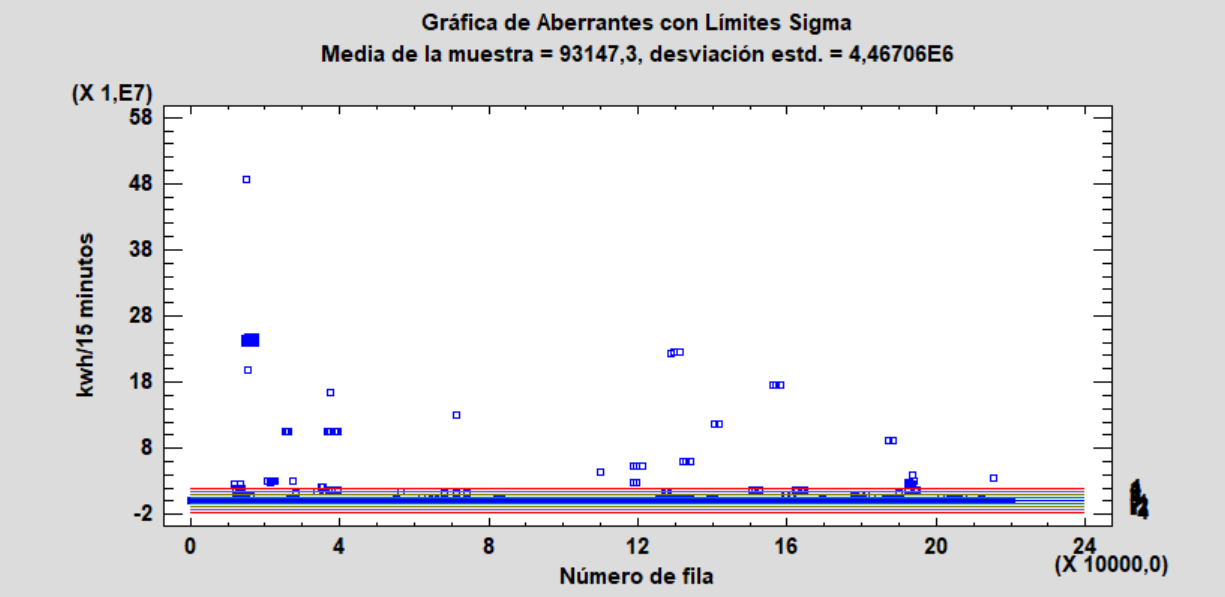
****

Consumo eléctrico general 5,37%

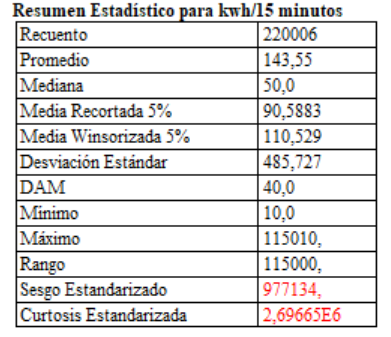
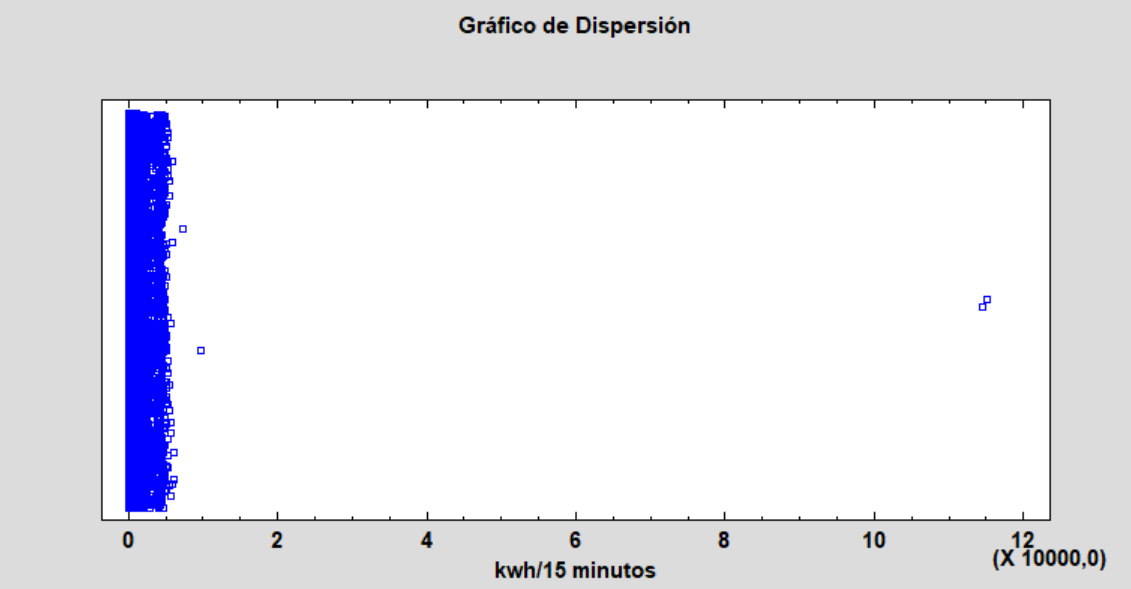
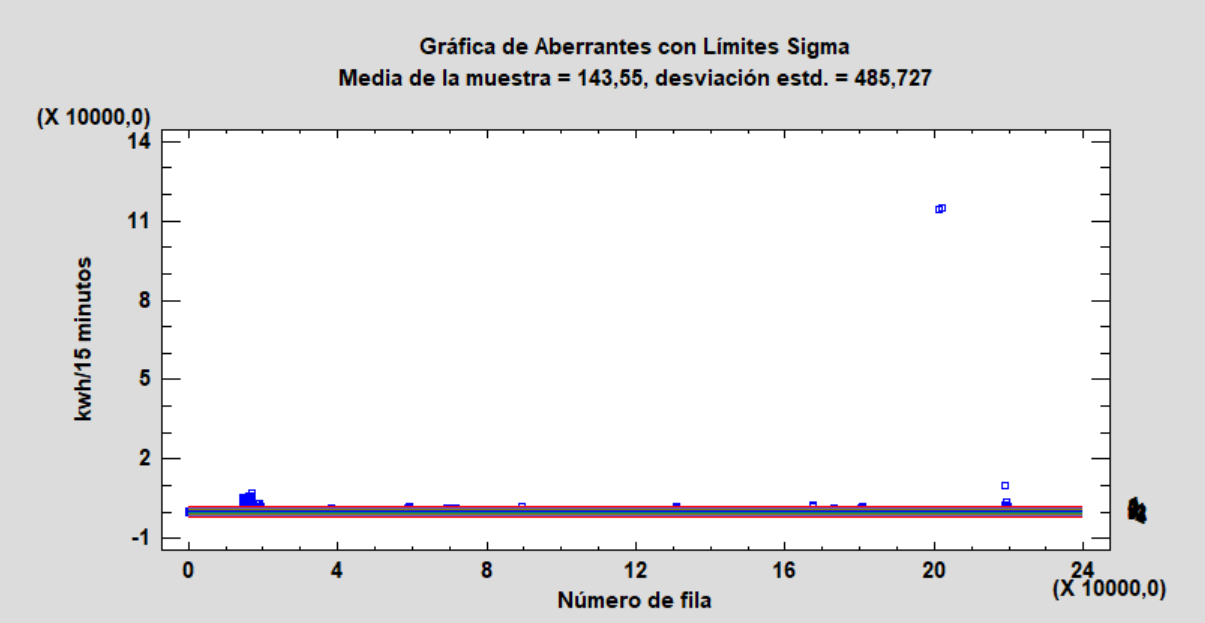
*Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M el 29,75%*

****

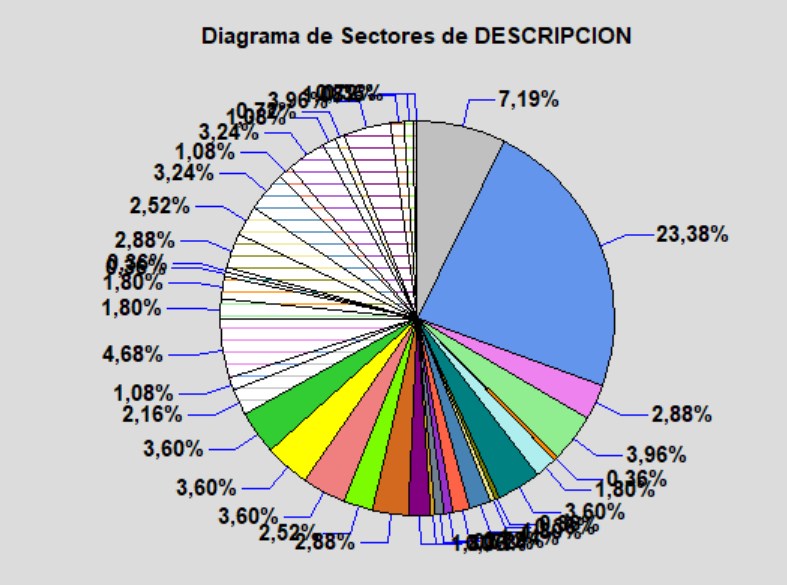
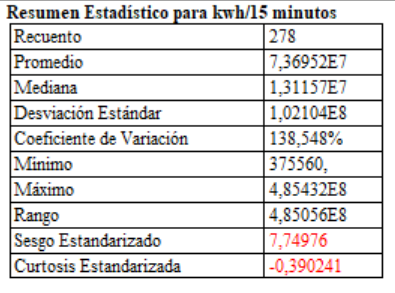
**Octubre**

****

**Después**

****

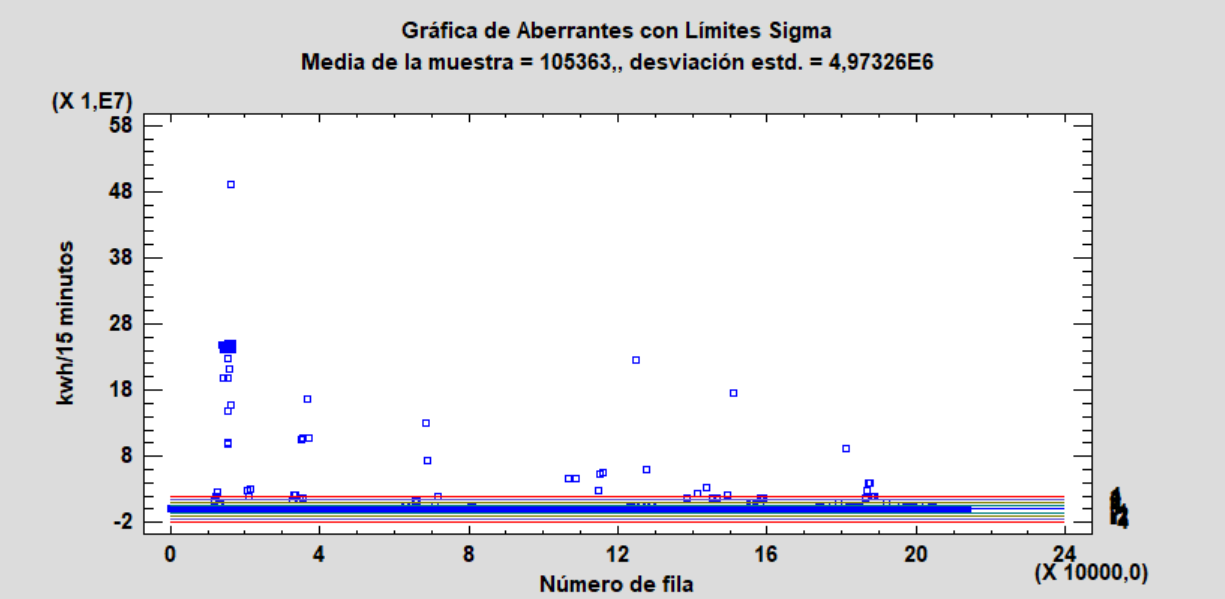
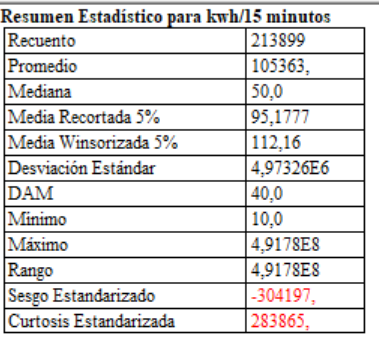
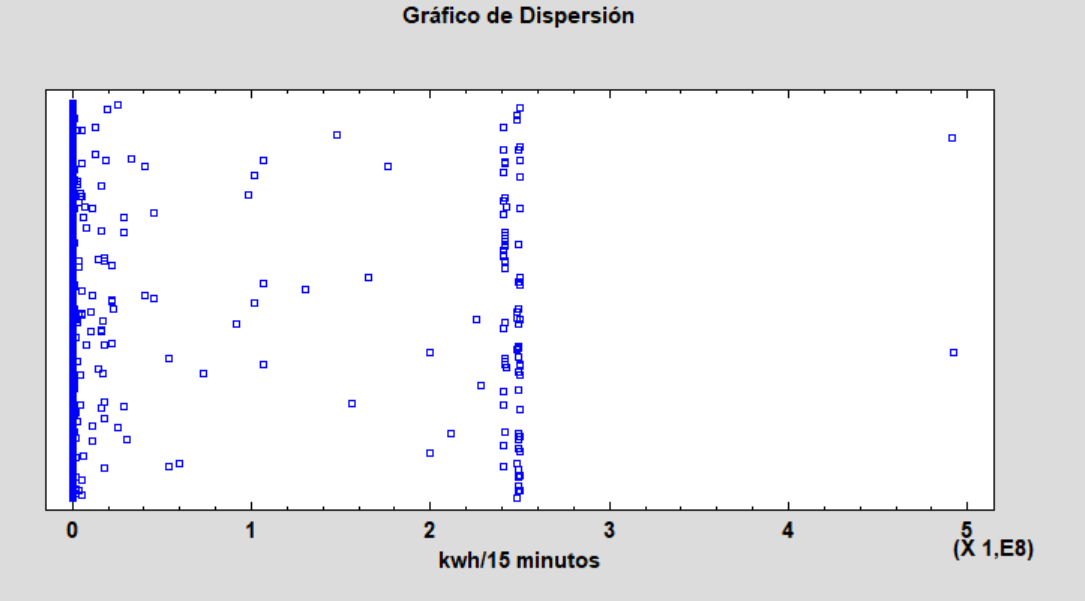
**Datos atípicos**

****

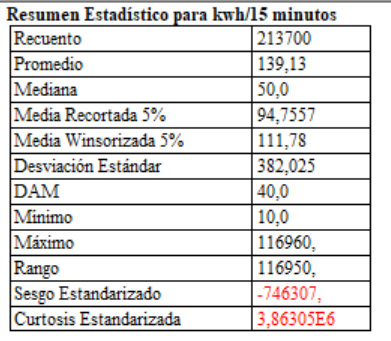
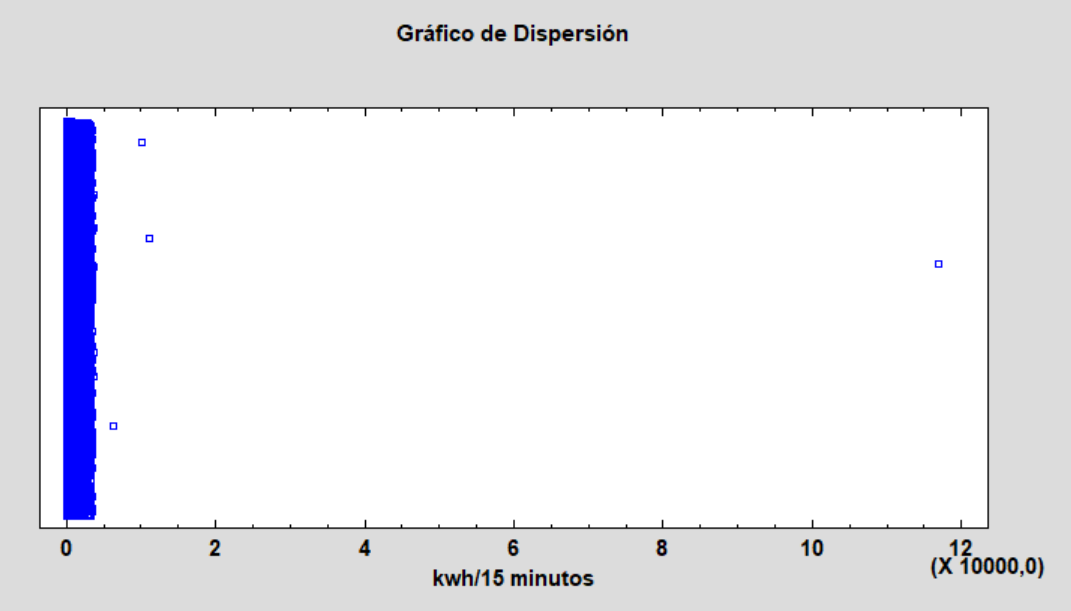
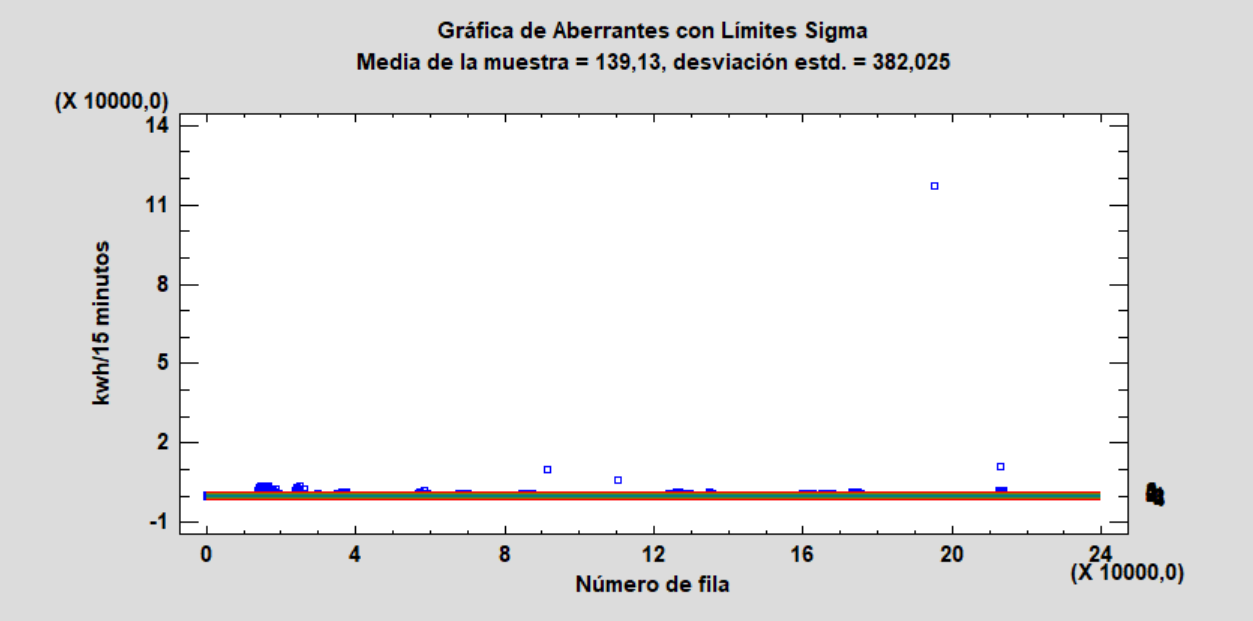
*Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M el 23,38%*

Consumo eléctrico general 7,19%

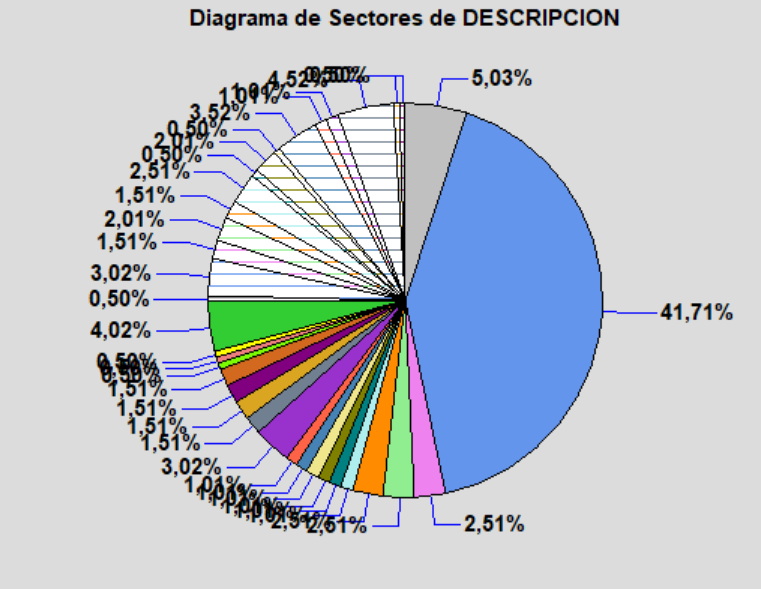
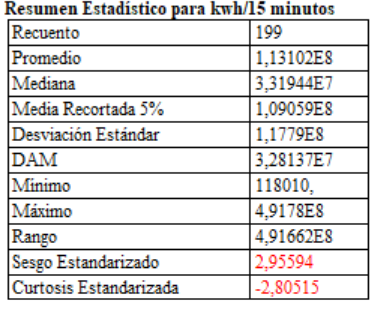
**Noviembre**

****

**después**

****

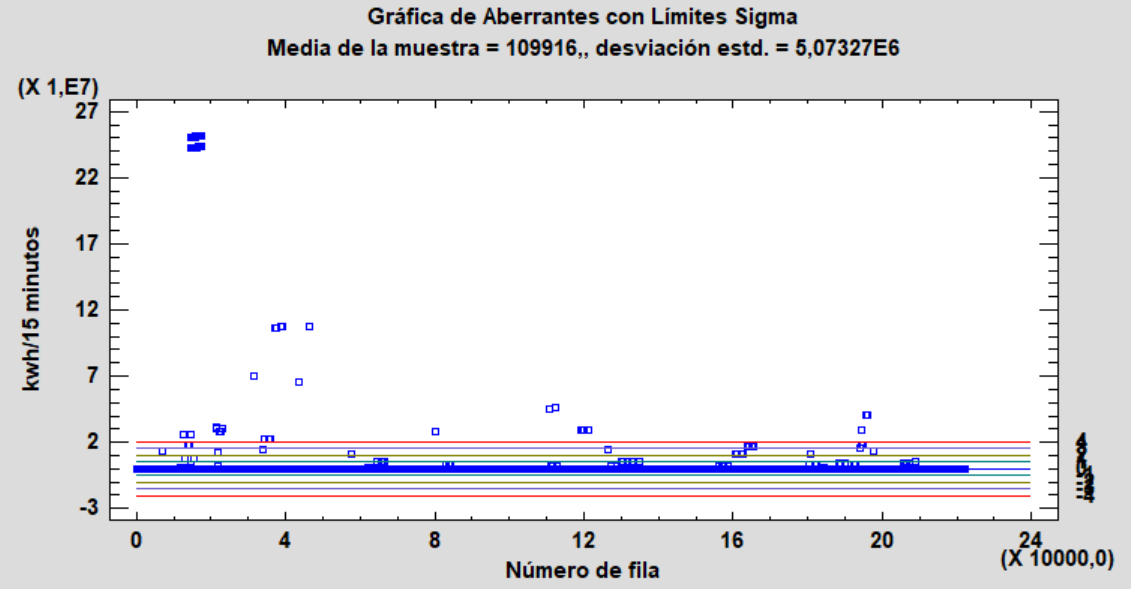
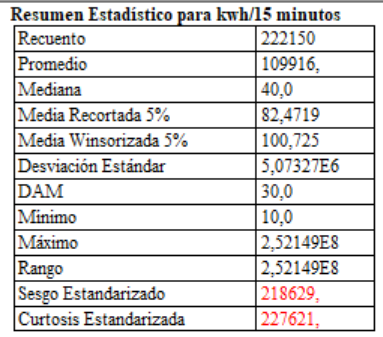
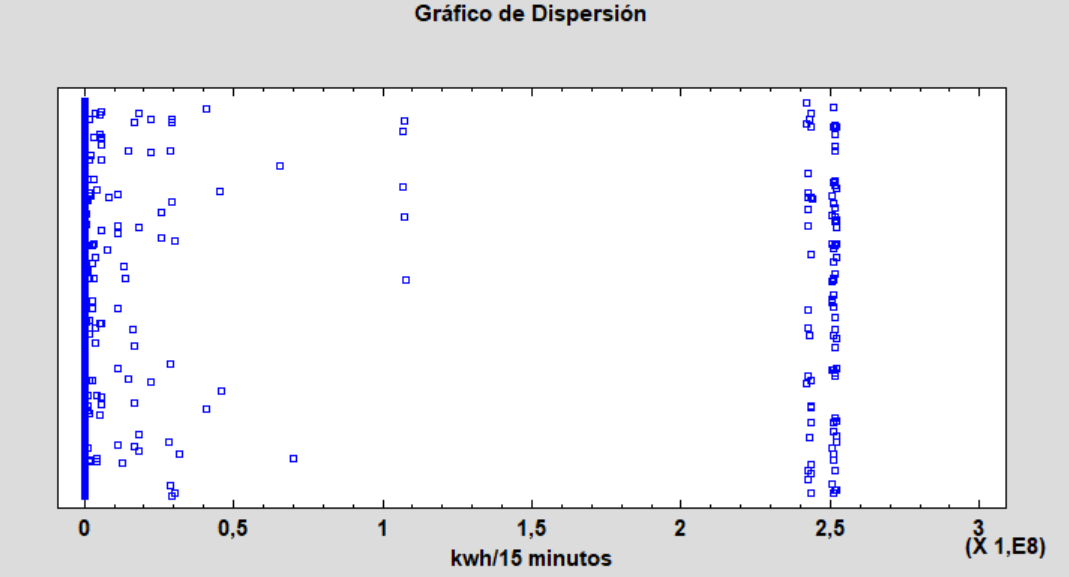
**Atípicos**

****

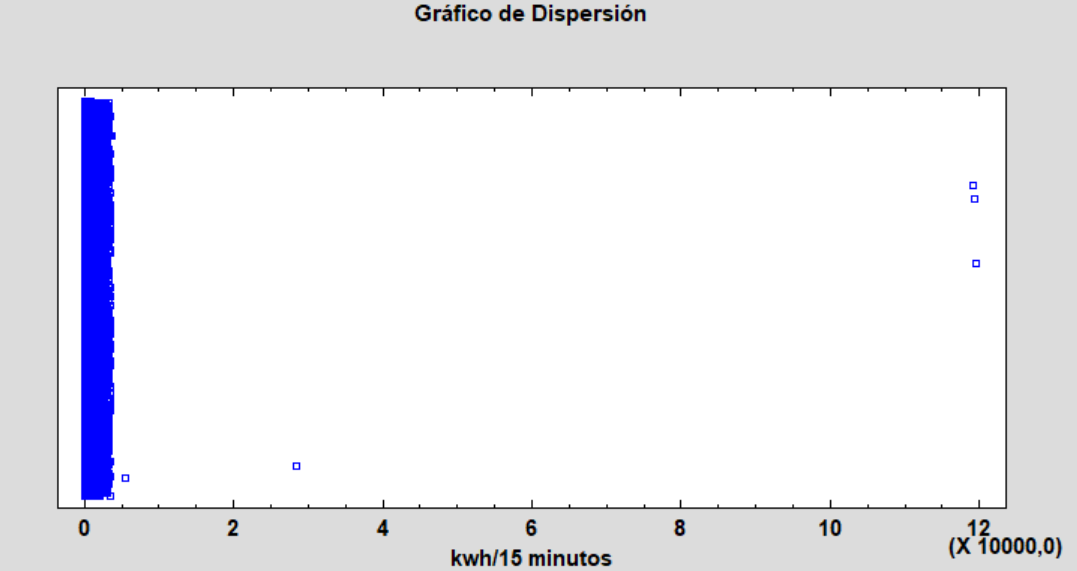
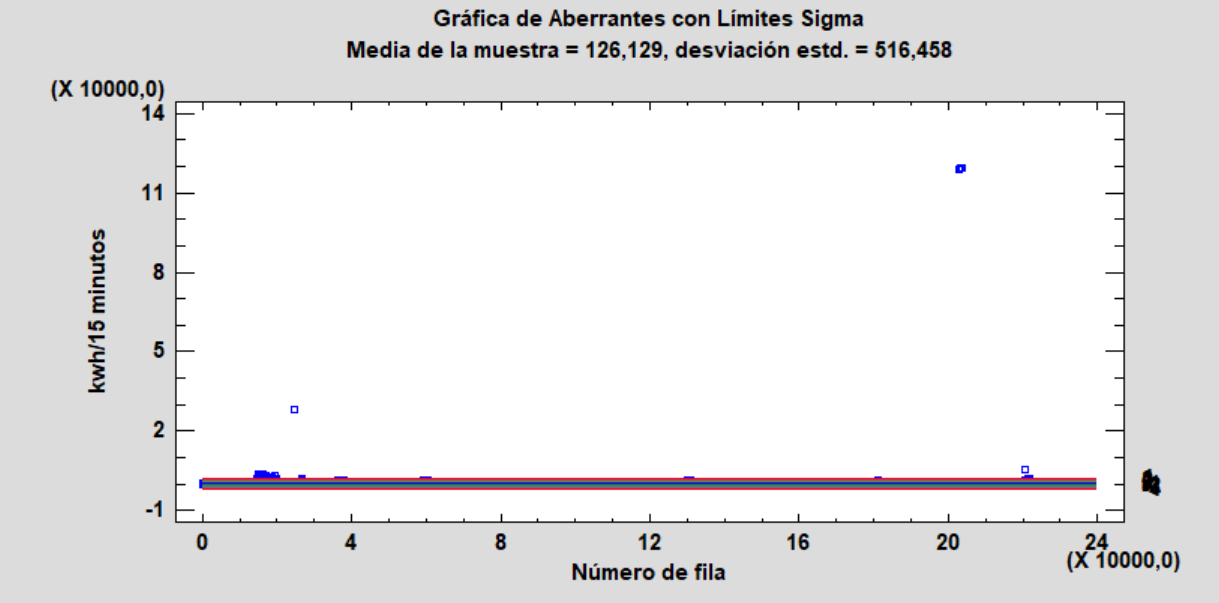
*Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M el 41,71%*

Consumo eléctrico general 5,03%

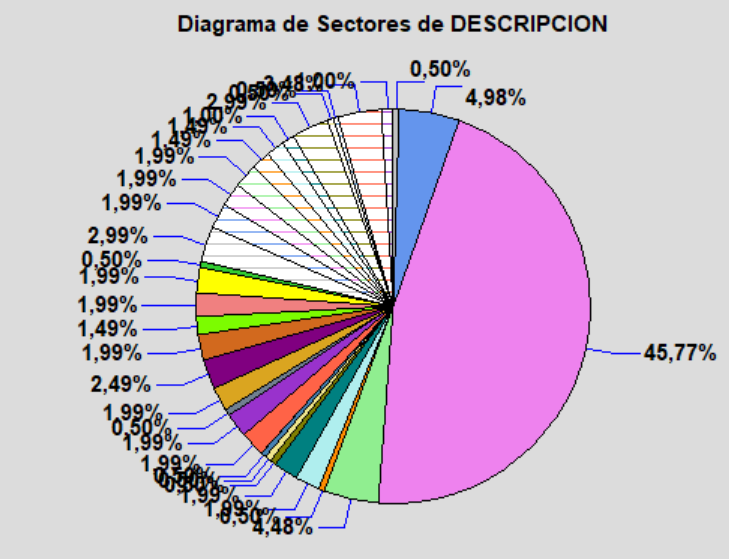
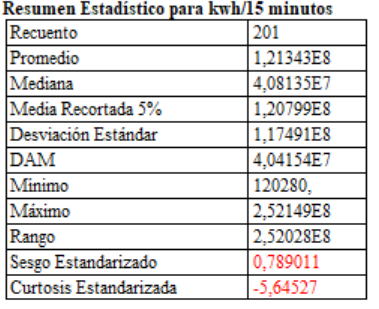
**Diciembre**

****

**Después**

****

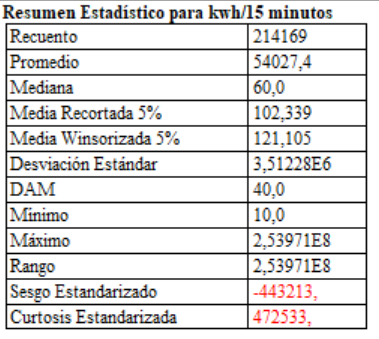
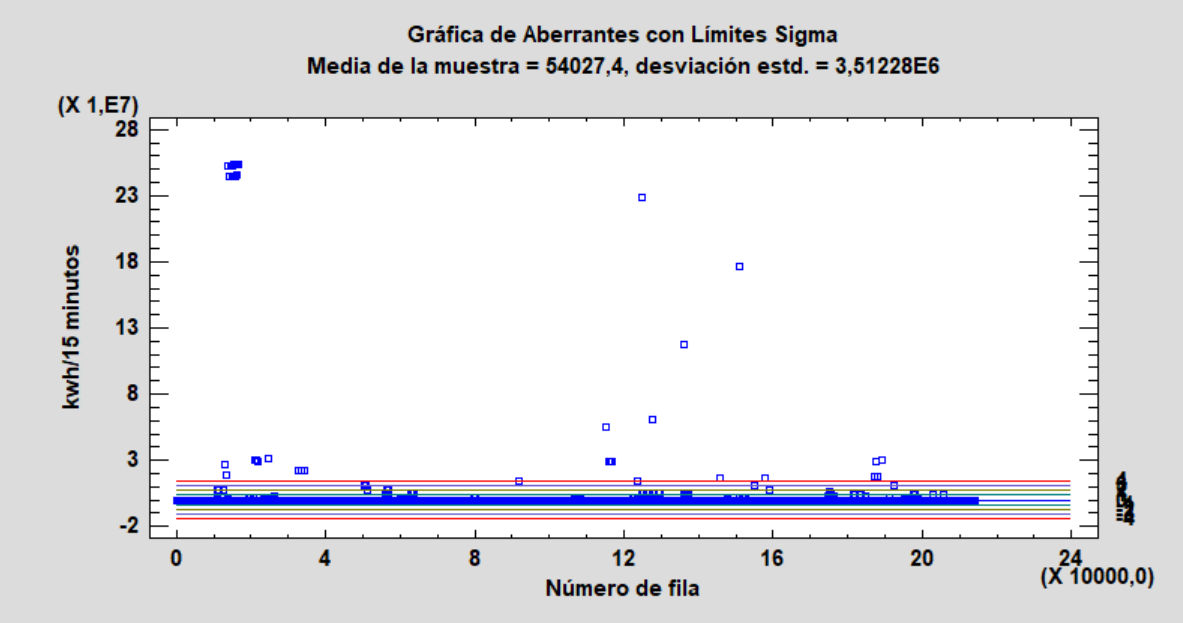
**Atípicos**

****

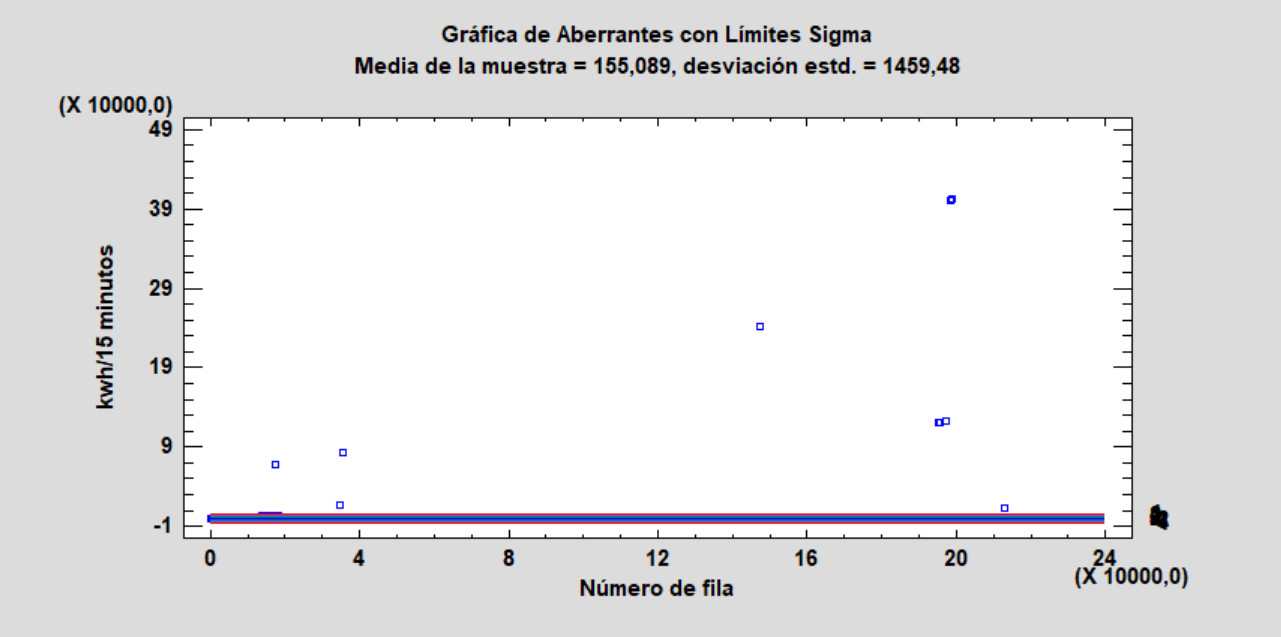
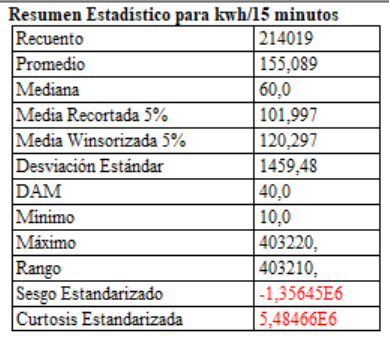
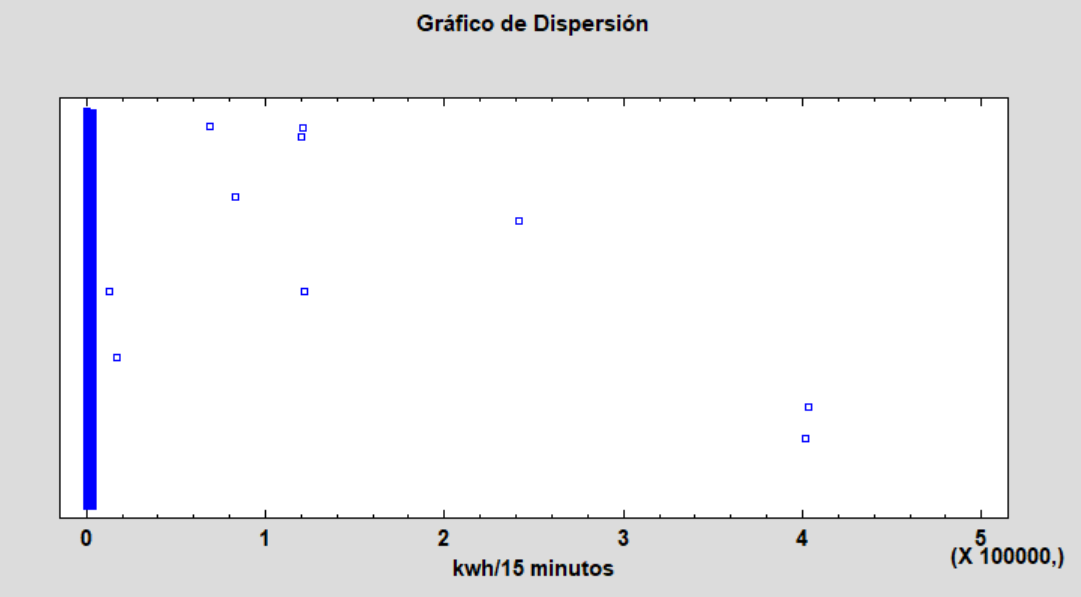
*Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M el 45,77%*

Consumo eléctrico general 4,98%

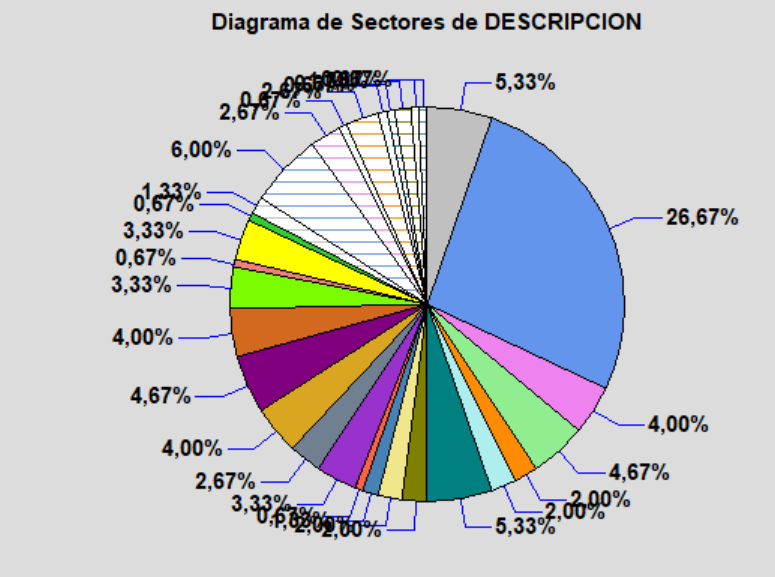
**Enero**

****

**Después**

****

**Atípicos**

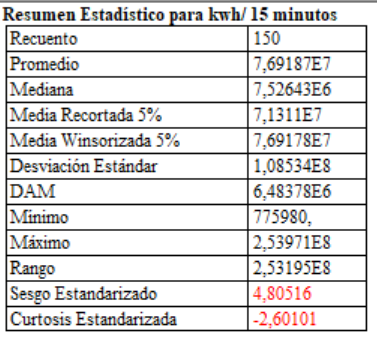
****

*Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M el 26,67%*

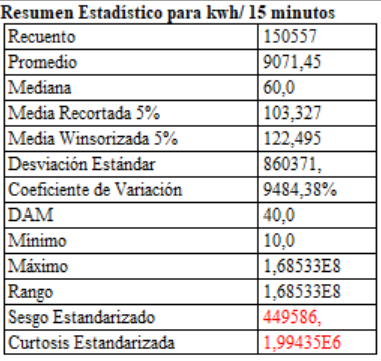
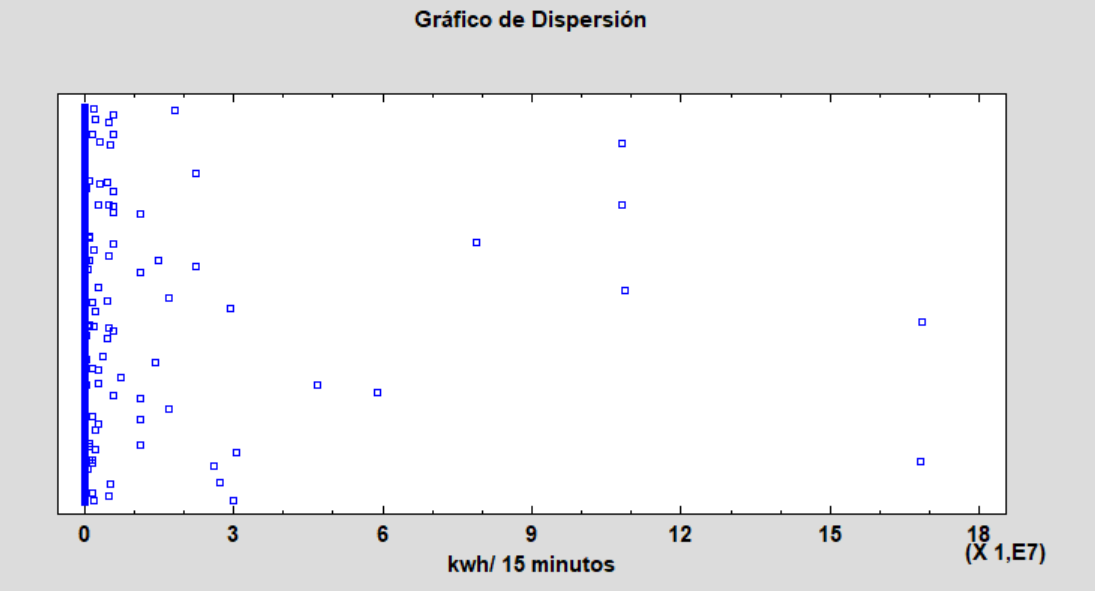
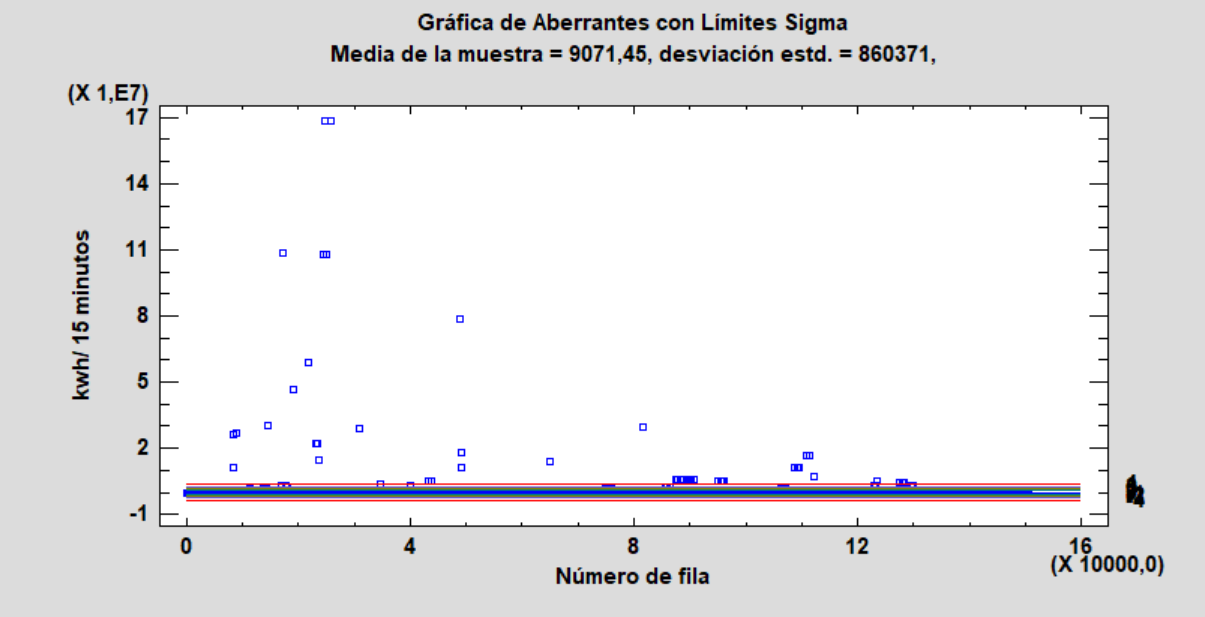
Consumo eléctrico general 5,33%

Consumo eléctrico general de los edificios: V1.A, V1.B y V1.D el 4,67%

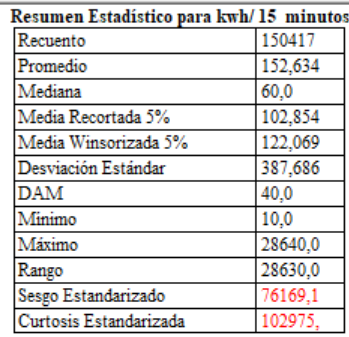
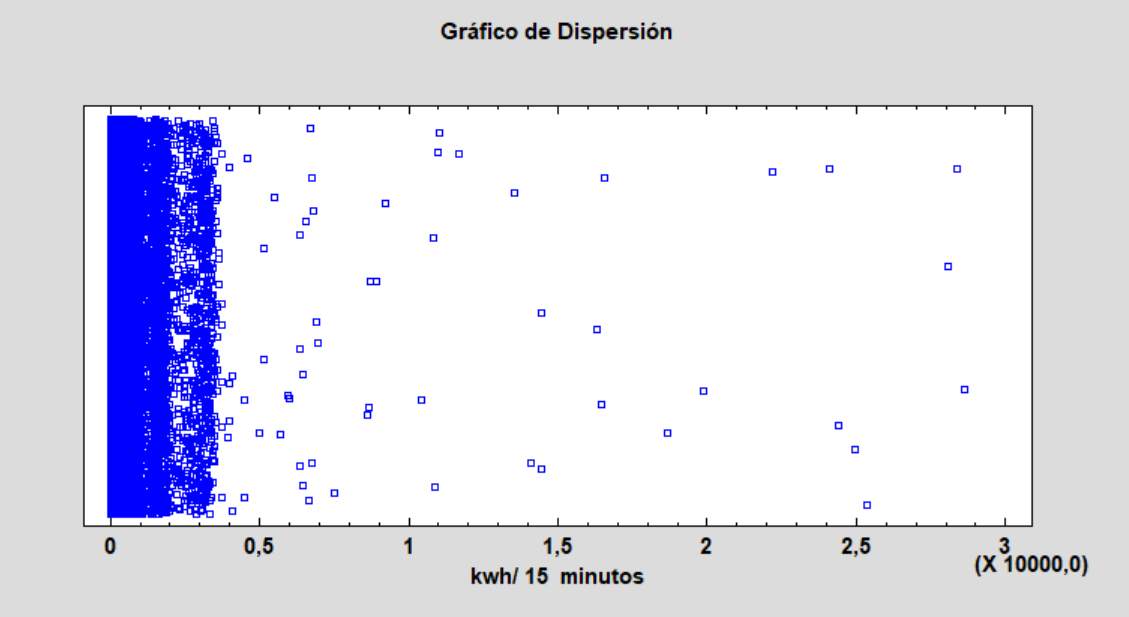
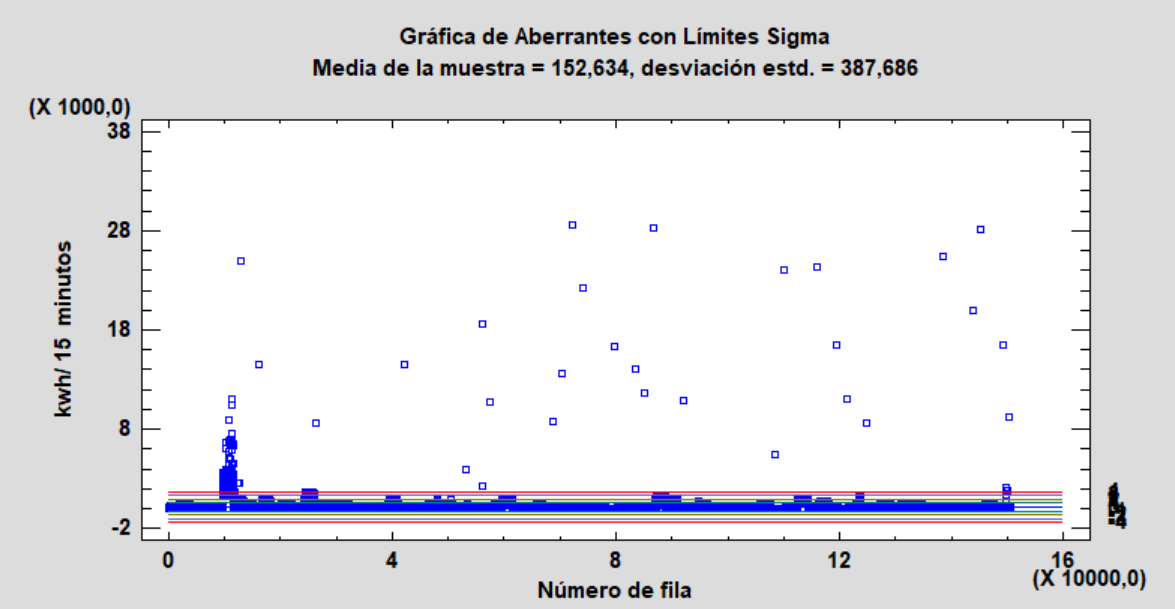
Consumo eléctrico general de los edificio: V8.P el 6%

****

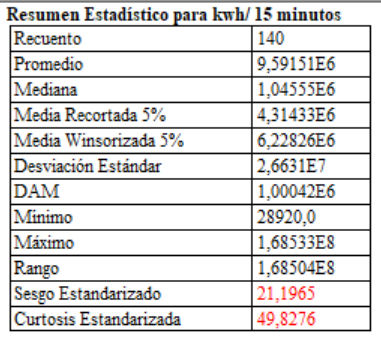
**Febrero**

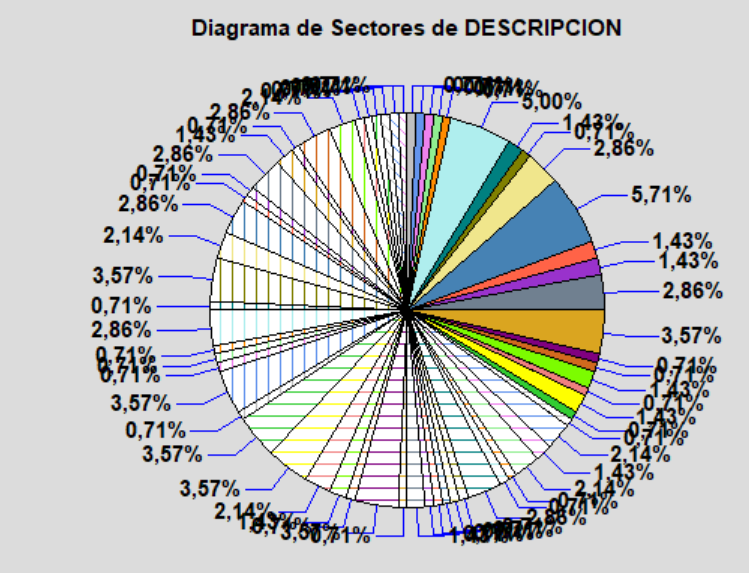
****

Después



Atípicos





*Consumo eléctrico general de la CPI: edificios V.8G, V.8E, V.8B, V.8M el 1,43%*

Consumo eléctrico general 5%

Consumo eléctrico general de los edificios: V1.A, V1.B y V1.D el 5,71%

**PARTE 2 (MARZO - AGOSTO) 2019**

En estos 6 meses hay muchos menos atípicos por mes, la forma de la que se ha procedido es:

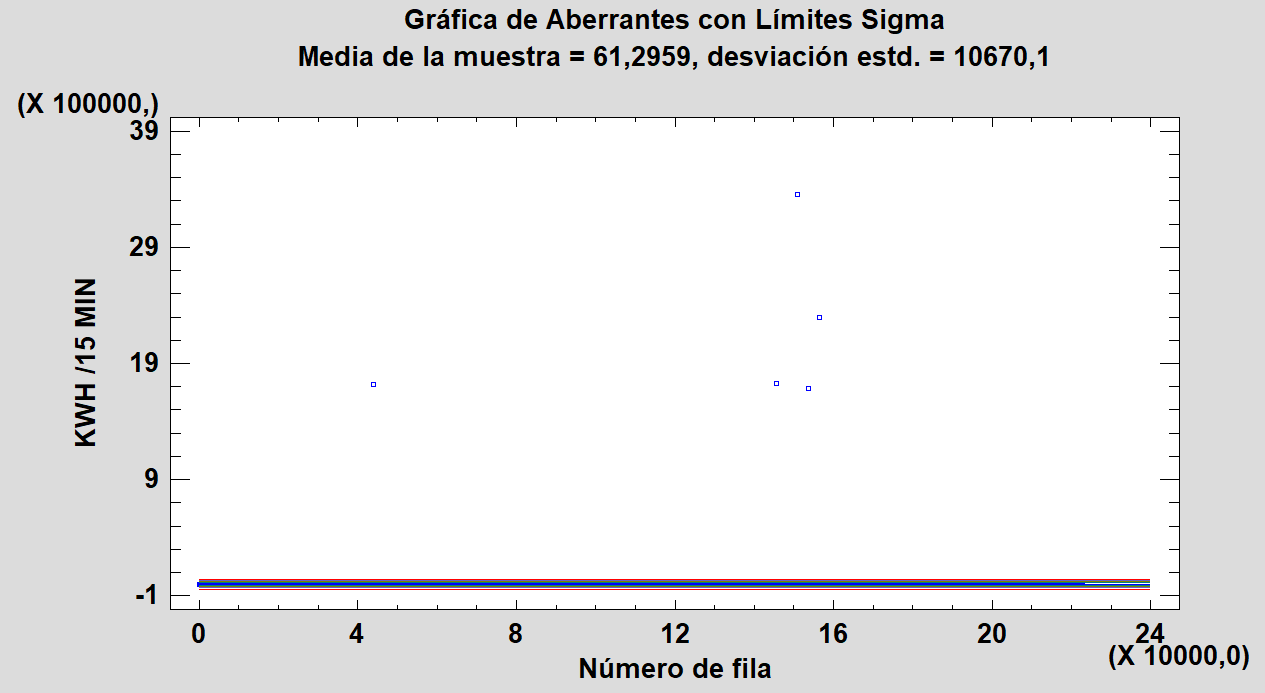
1º Se ha identificado los datos atípicos de KWH/15 MIN mediante un gráfico de aberrantes en el programa Statgraphics

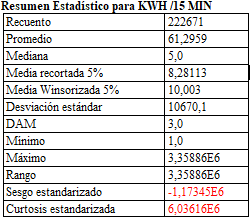
2º Se ha confirmado la existencia de estos datos buscando las filas en el excel del mes correspondiente.

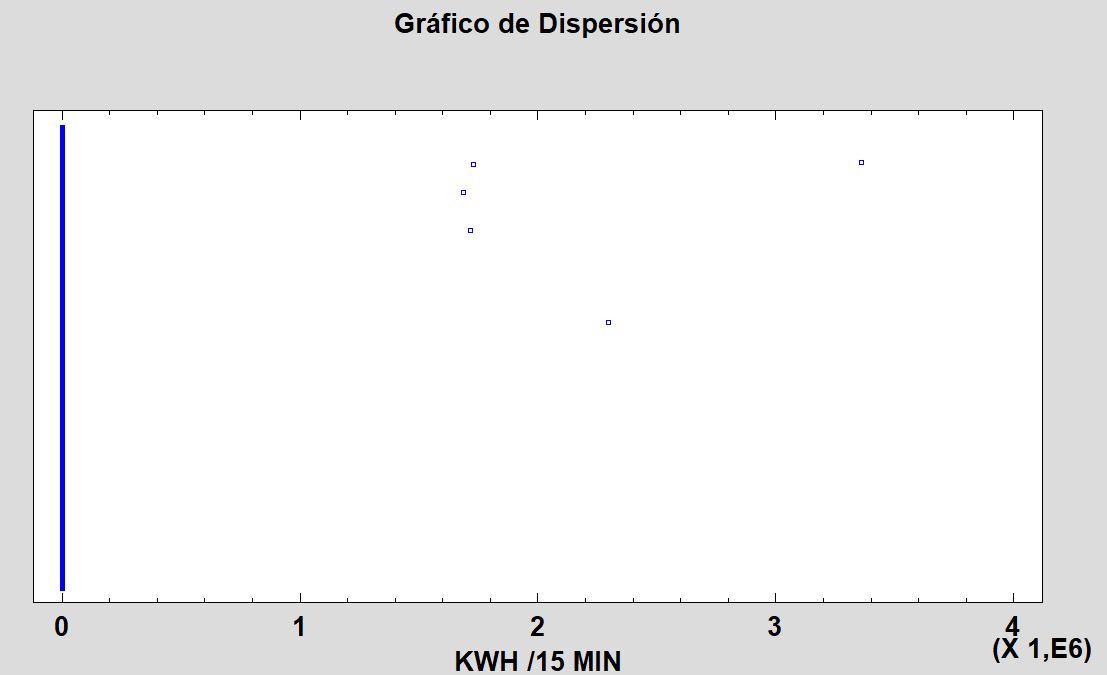
3º Dada la poca cantidad de datos atípicos, para eliminarlos se borraron directamente del excel, buscándolos uno a uno.

***MARZO:***

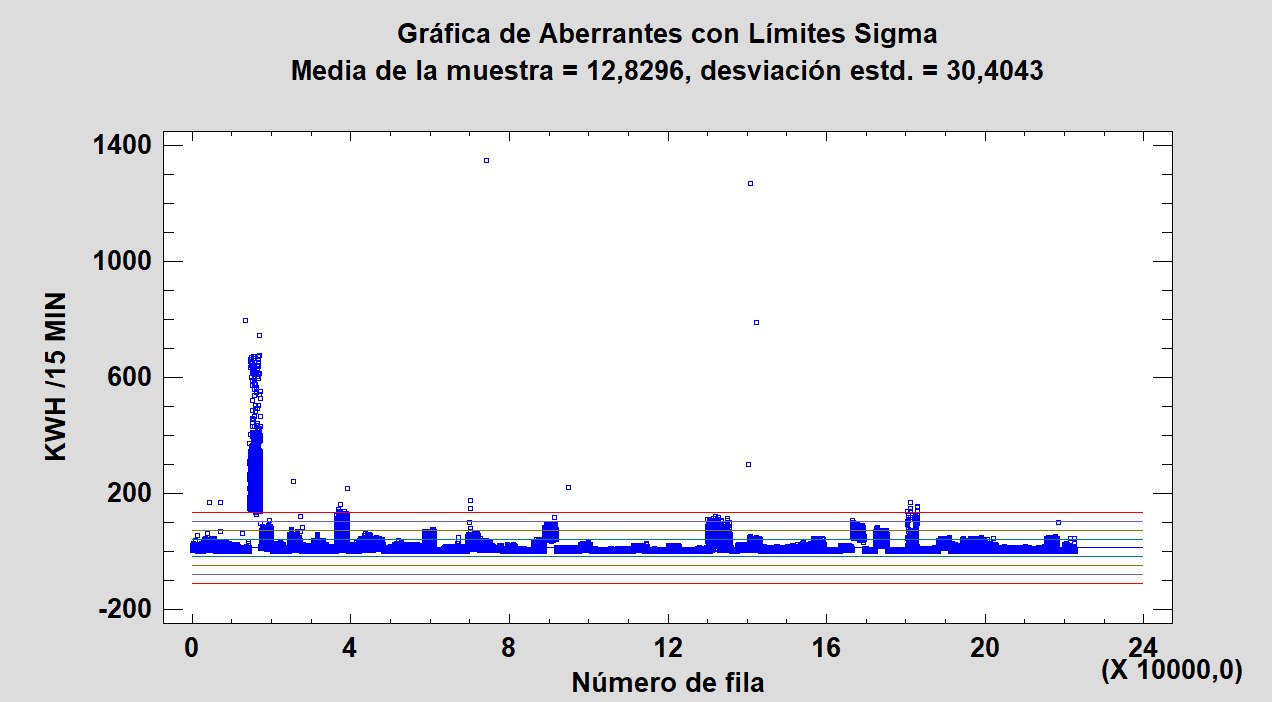
Datos antes de la limpieza:

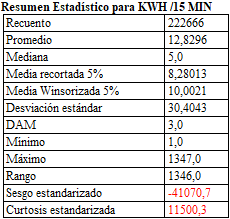


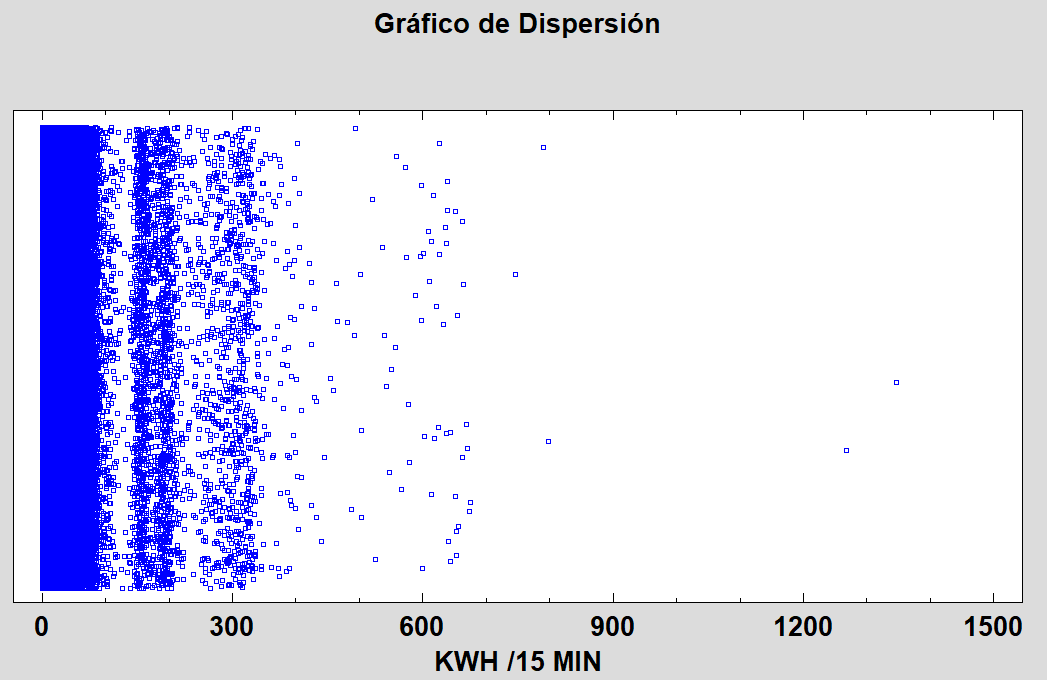




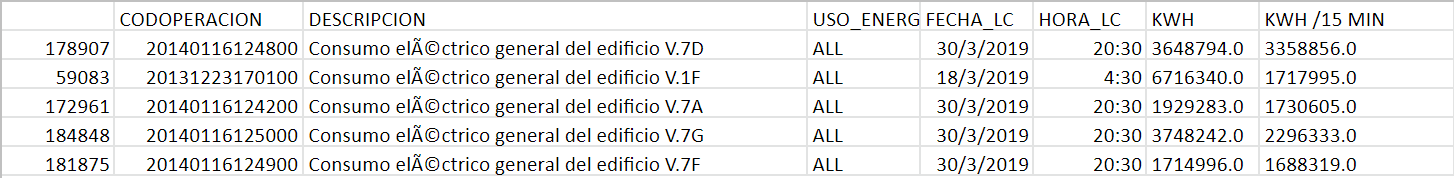
Datos después de la limpieza





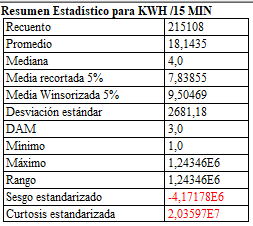
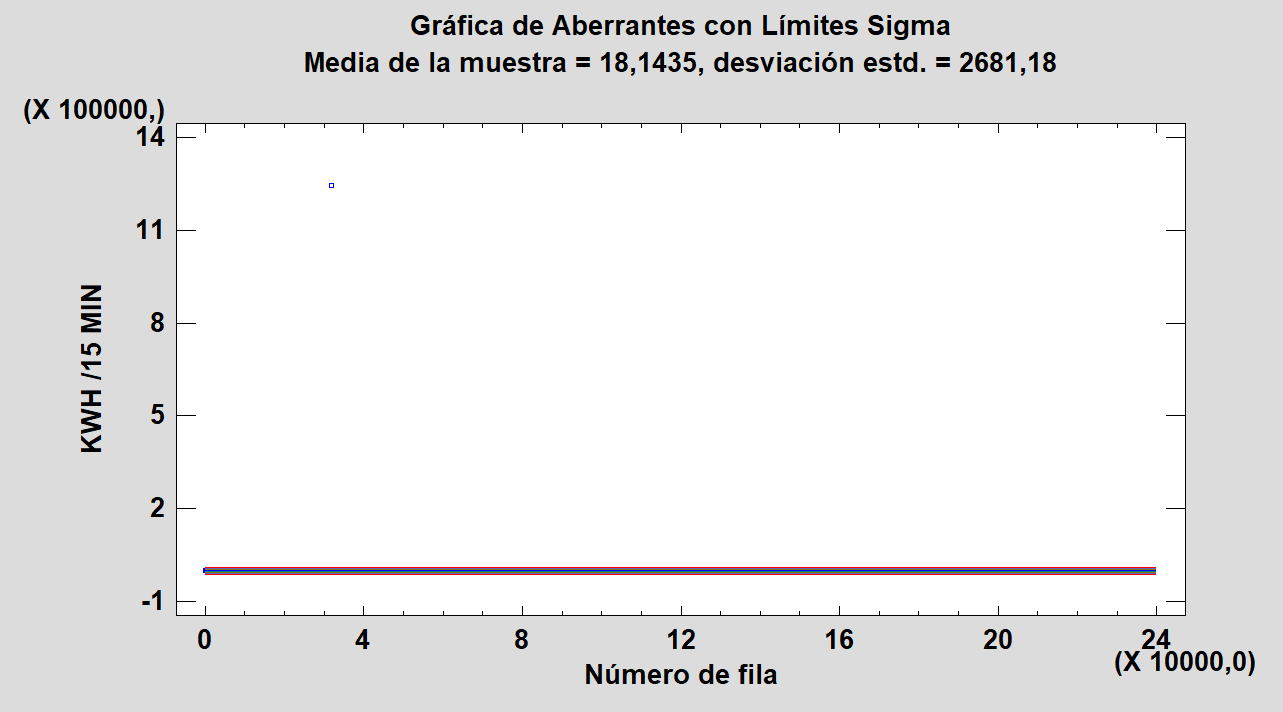


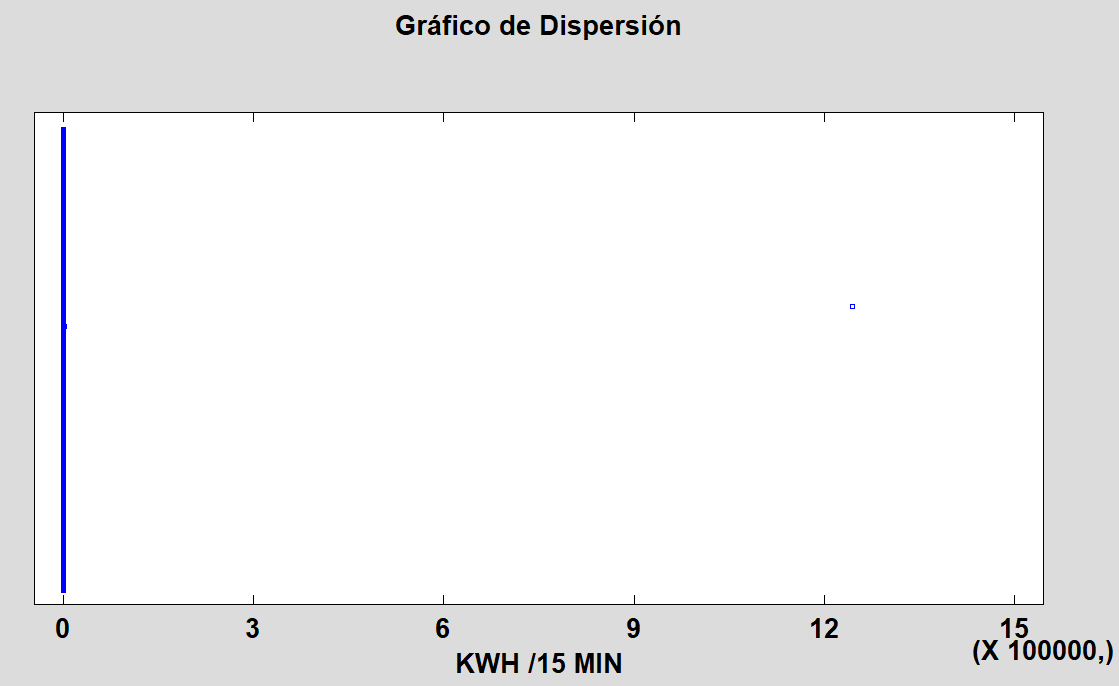
Estos han sido los datos eliminados en Marzo:

Como podemos ver, la media y la desviación típica se han reducido en 50 y 10600 puntos, mientras que los parámetros robustos se han mantenido igual. Los valores eliminados son tan altos que eclipsan a todos los demás en el gráfico de dispersión y de aberrantes.

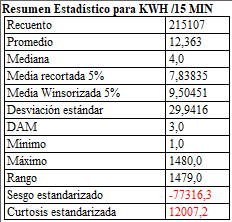
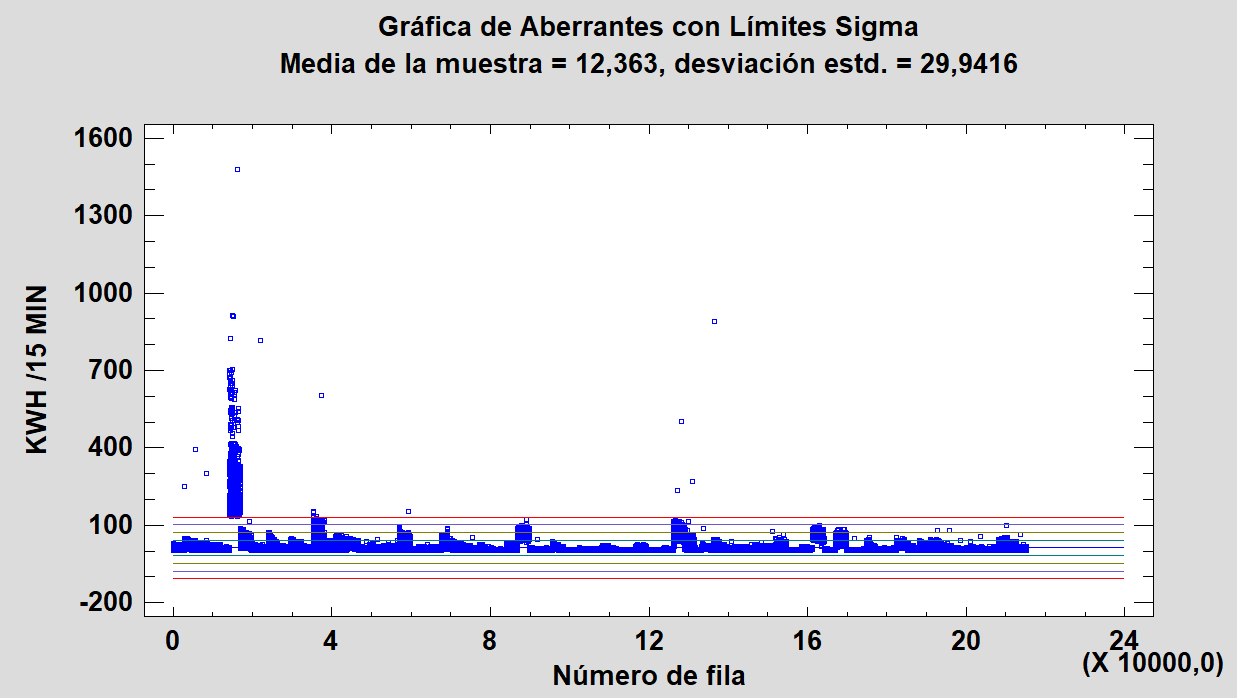
***ABRIL:***

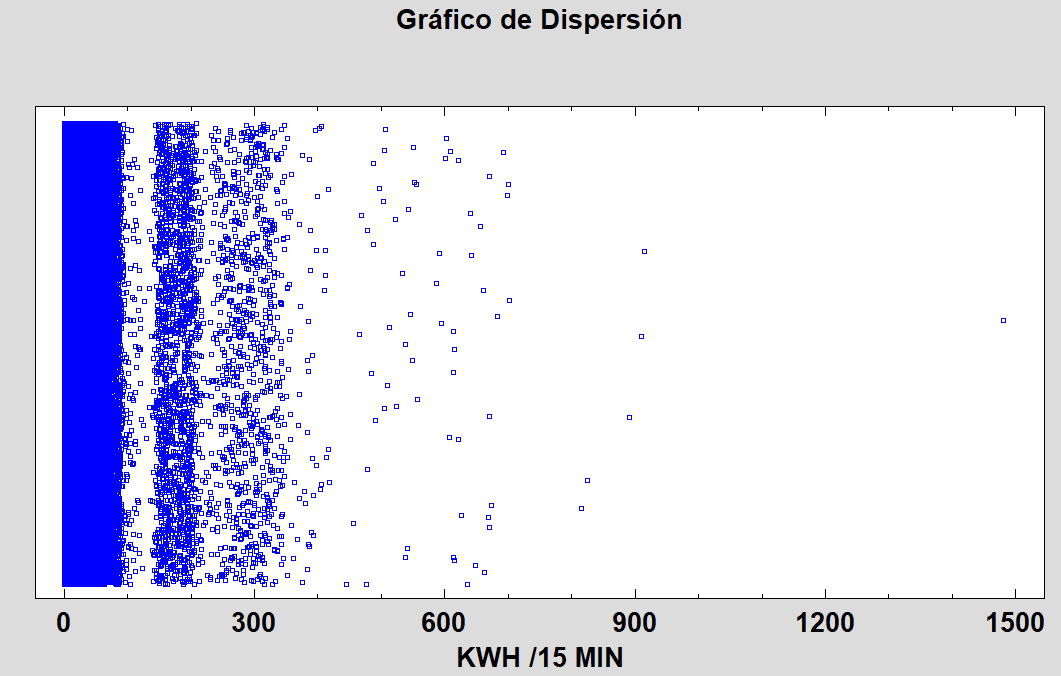
Datos antes de la limpieza:





Datos después de la limpieza:





Este ha sido el dato eliminado en Abril:

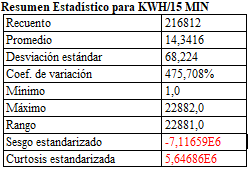


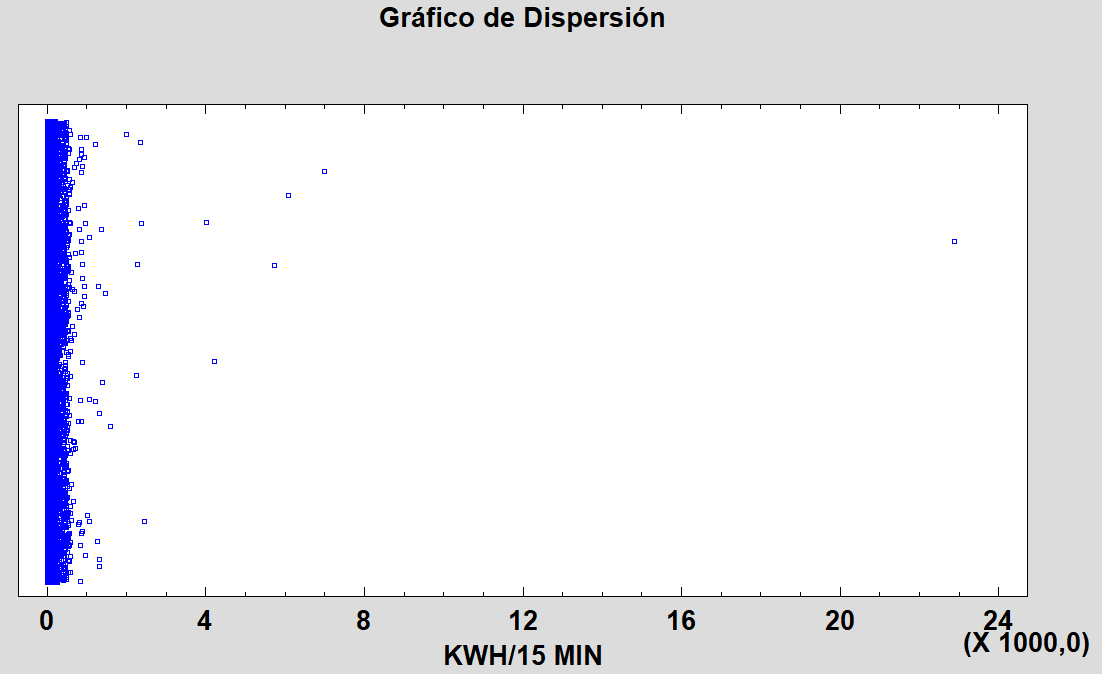
Como podemos ver, un sólo ha dato ha logrado bajar la media 6 puntos y la desviación típica 2650 puntos, mientras que los parámetros robustos se han mantenido igual. El valor eliminado es tan alto que eclipsa a todos los demás en el gráfico de dispersión y de aberrantes.

***MAYO:***

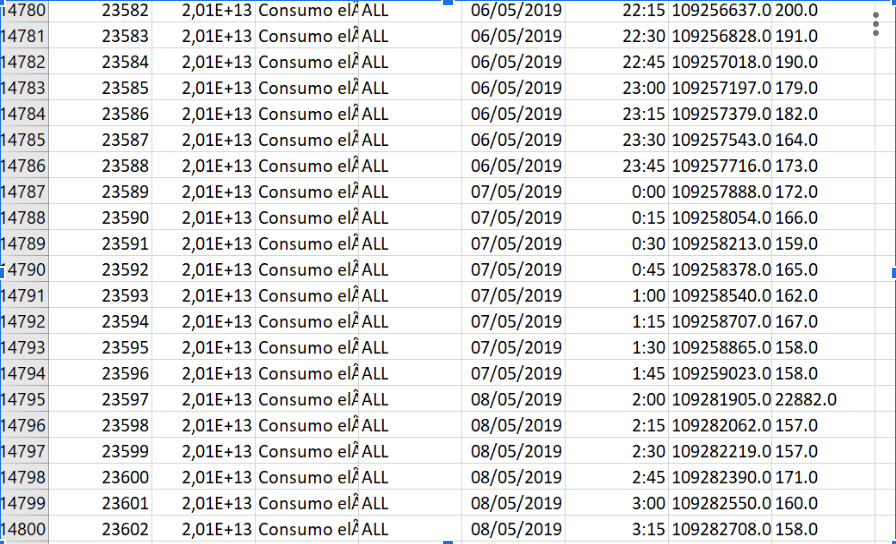
*En este mes ningún dato ha sido eliminado*

Datos:



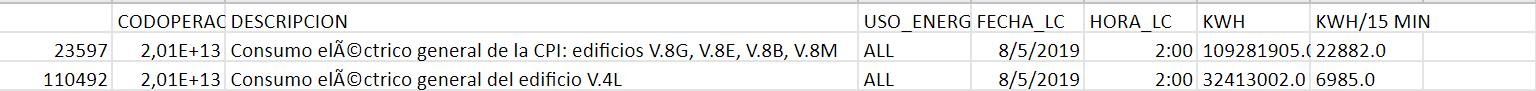


En este mes se puede ver un ejemplo de valores acumulados en un sólo dato, por la razón que hemos explicado más arriba, especialmente el dato de 22800 KWH/15min, así ha ocurrido:

*Fila 14794 -> 14795*

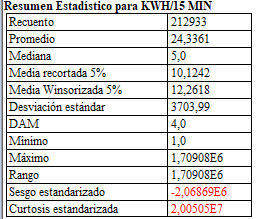
Aquí podemos ver como un edificio que gasta mucho más que lo habitual (~15) cada 15 minutos (+150), deja de registrar datos durante un día, lo cual implica que faltan 15 \* 4 \* 24 = 1440 datos, cada uno de ellos de alrededor de 150 KW, dan lugar a una acumulación de 22800 KW.

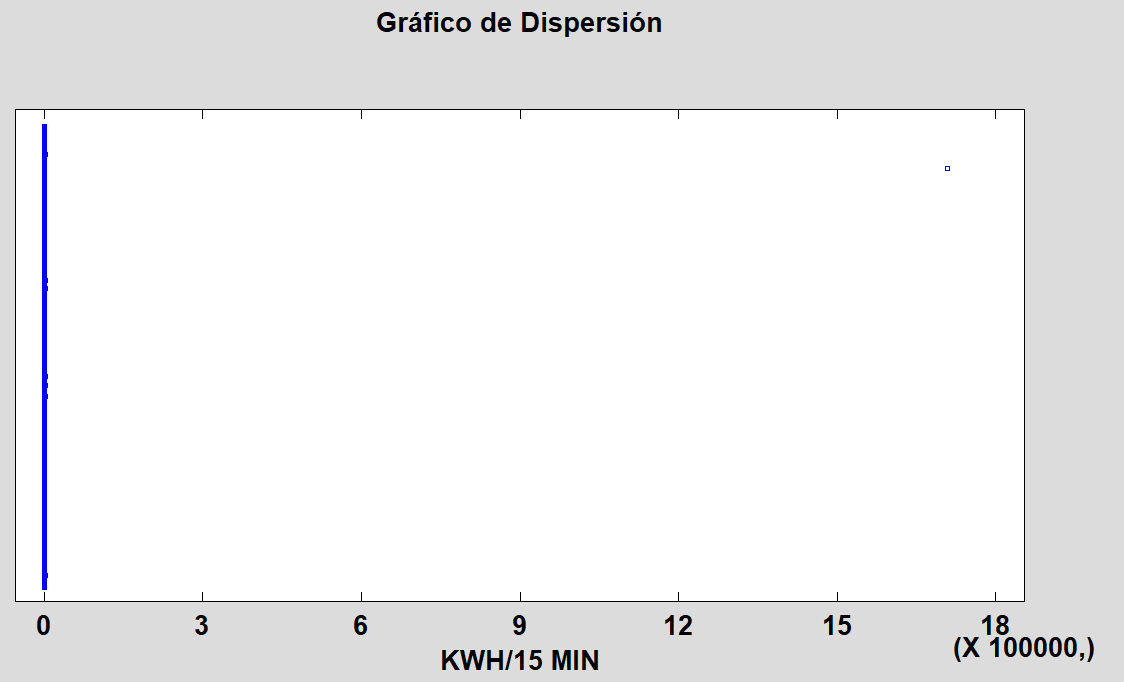
Estos son algunos de los datos extremos de este mes:



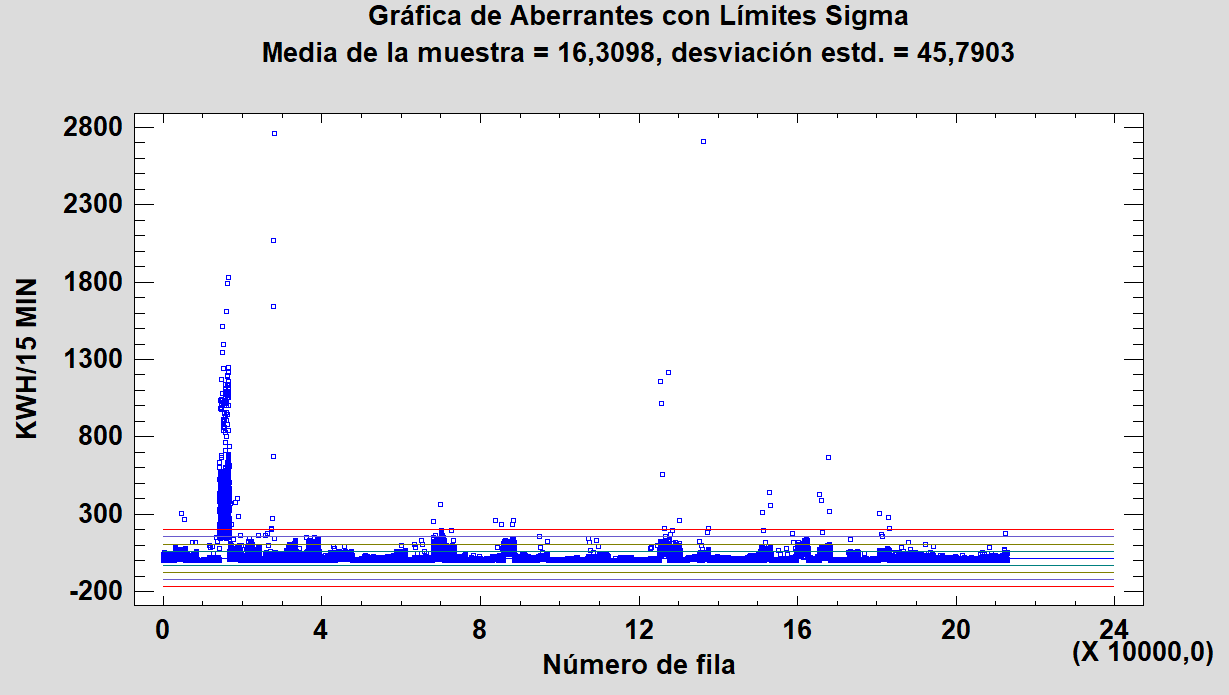
***JUNIO:***

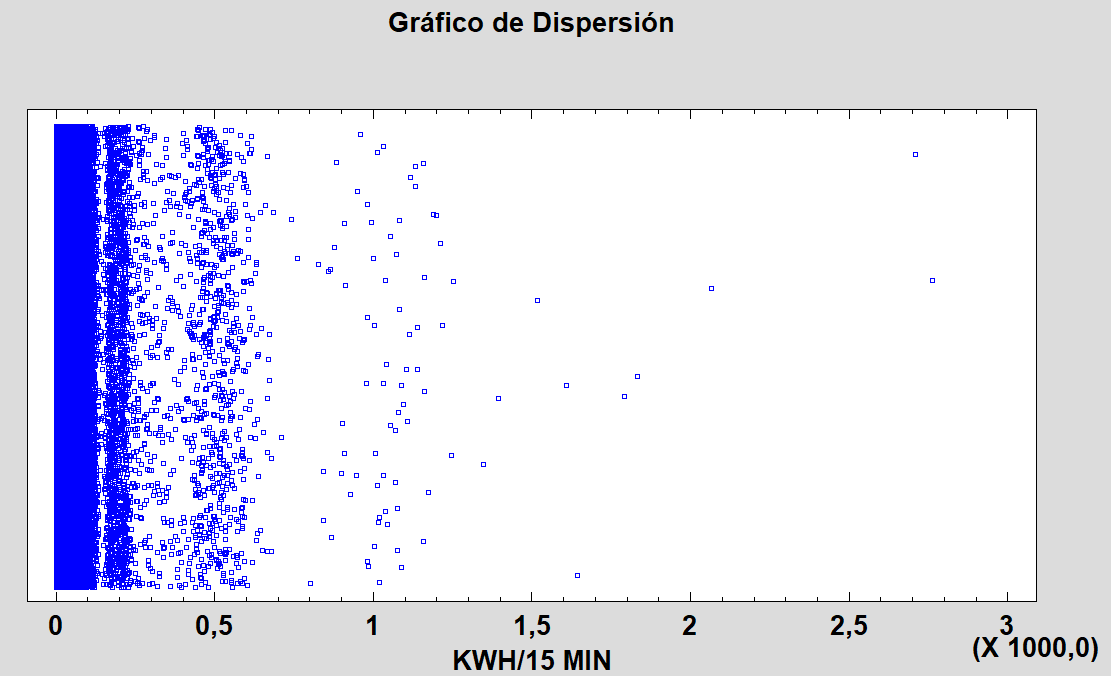
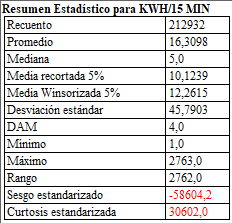
Datos antes de la limpieza:





Datos después de la limpieza:





Este ha sido el dato eliminado en Junio:

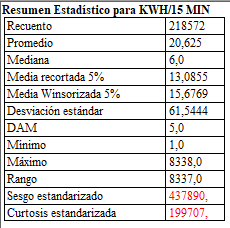
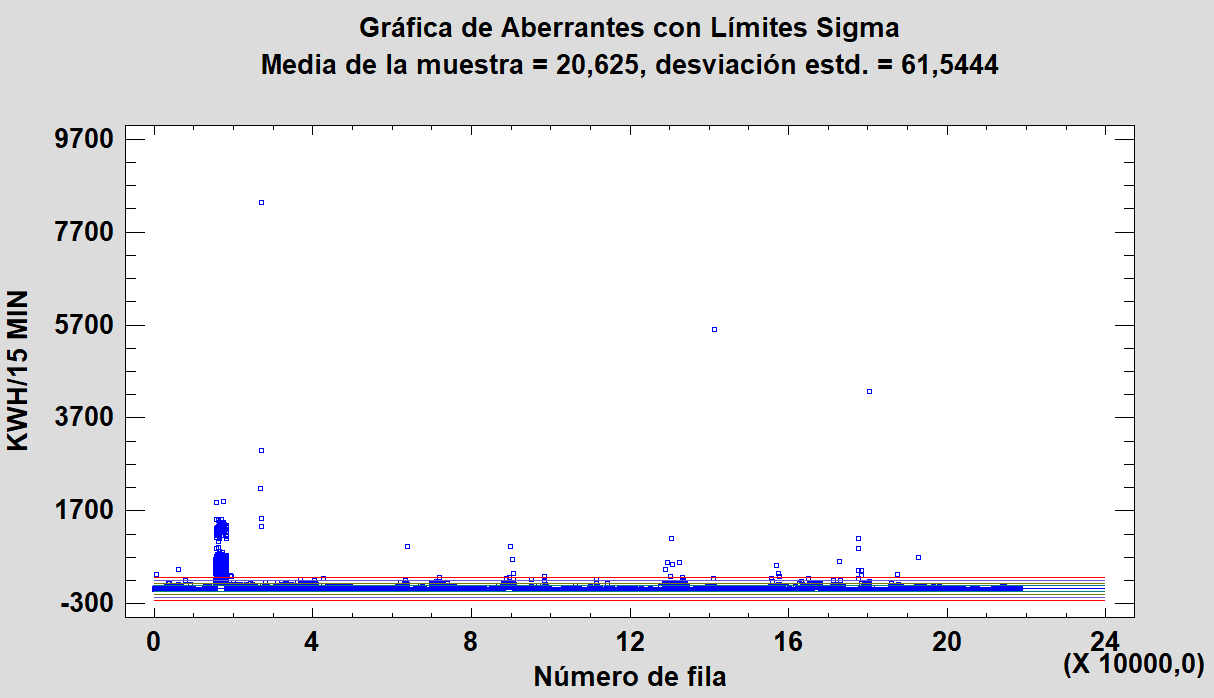


Como podemos ver, un sólo ha dato ha logrado bajar la media 8 puntos y la desviación típica 3650 puntos, mientras que los parámetros robustos se han mantenido igual. El valor eliminado es tan alto que eclipsa a todos los demás en el gráfico de dispersión y de aberrantes.

***JULIO:***

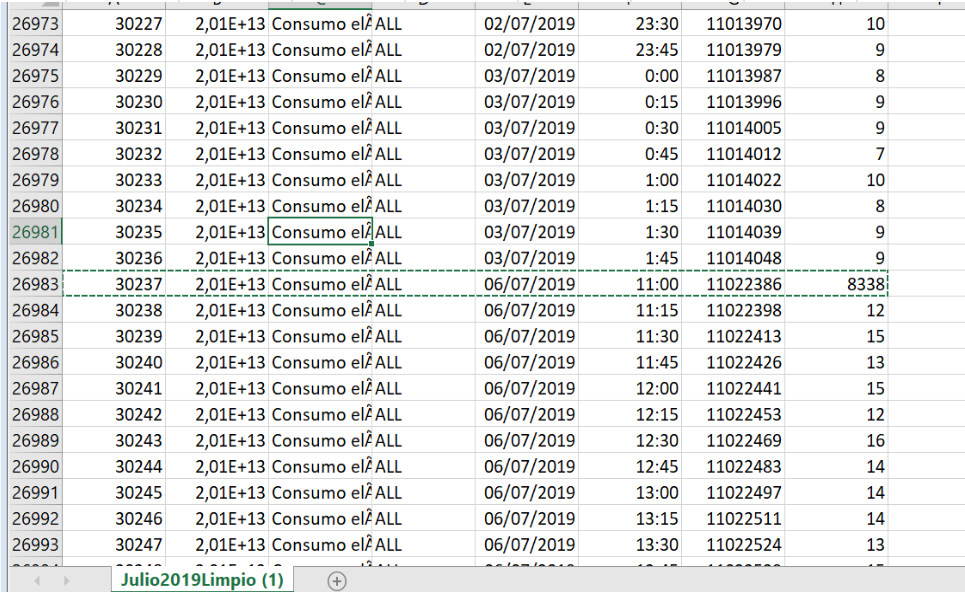
*En este mes ningún dato ha sido eliminado*

Datos:



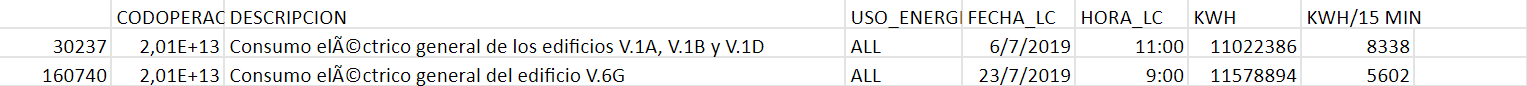
******

En este mes se puede ver un ejemplo de valores acumulados, al igual que en mayo, en un sólo dato, por la razón que hemos explicado más arriba, por ejemplo el dato más alto (8000 kwh/15min) se ha formado así:

*Fila 26982 -> 26983*

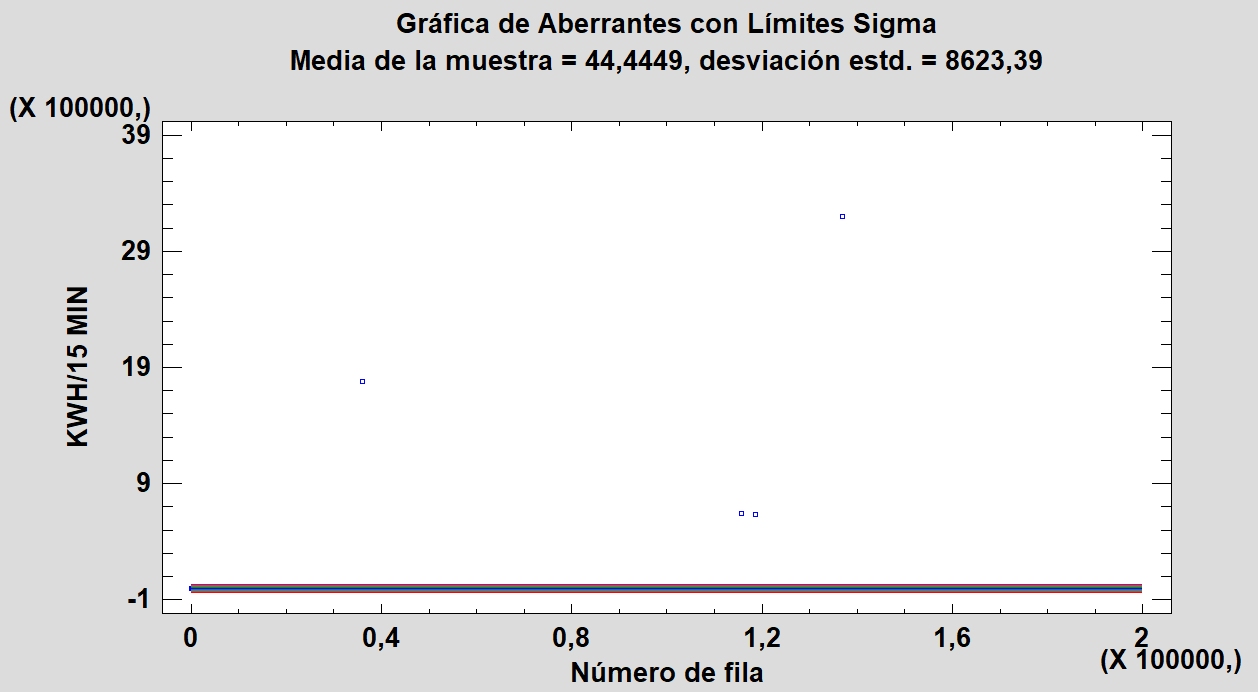
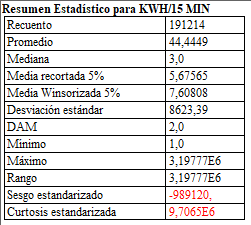
Aquí hay un fallo durante varios días, sin embargo los valores de este edificio no son tan altos como los del anterior y sólo se acumulan 8000 kw. Faltan unos 15\*4 \* 81 = 4860 datos

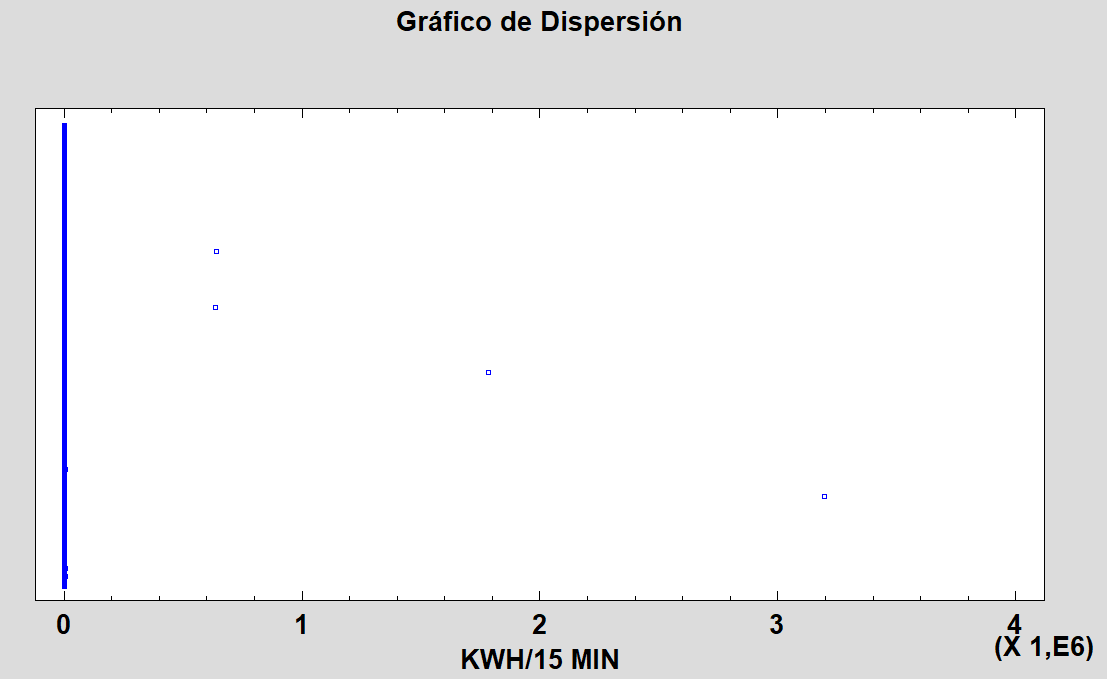
Estos son algunos de los datos extremos de este mes:



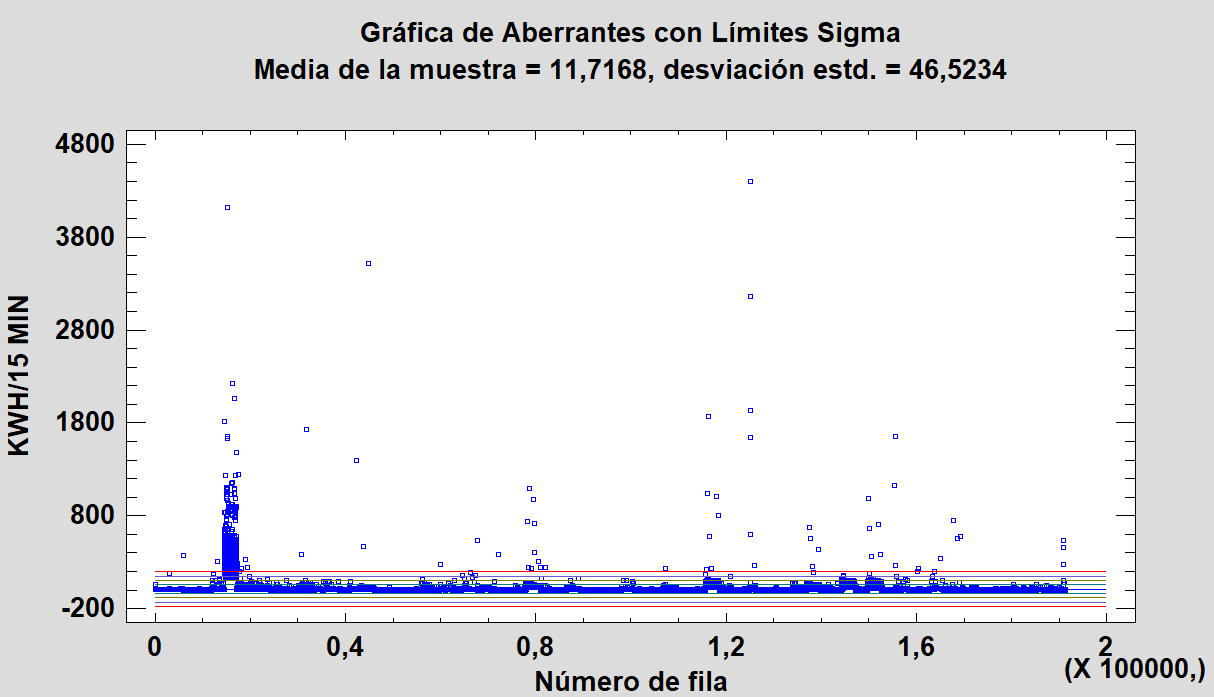
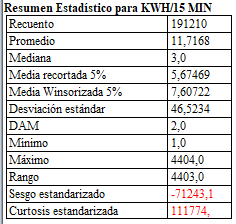
***AGOSTO:***

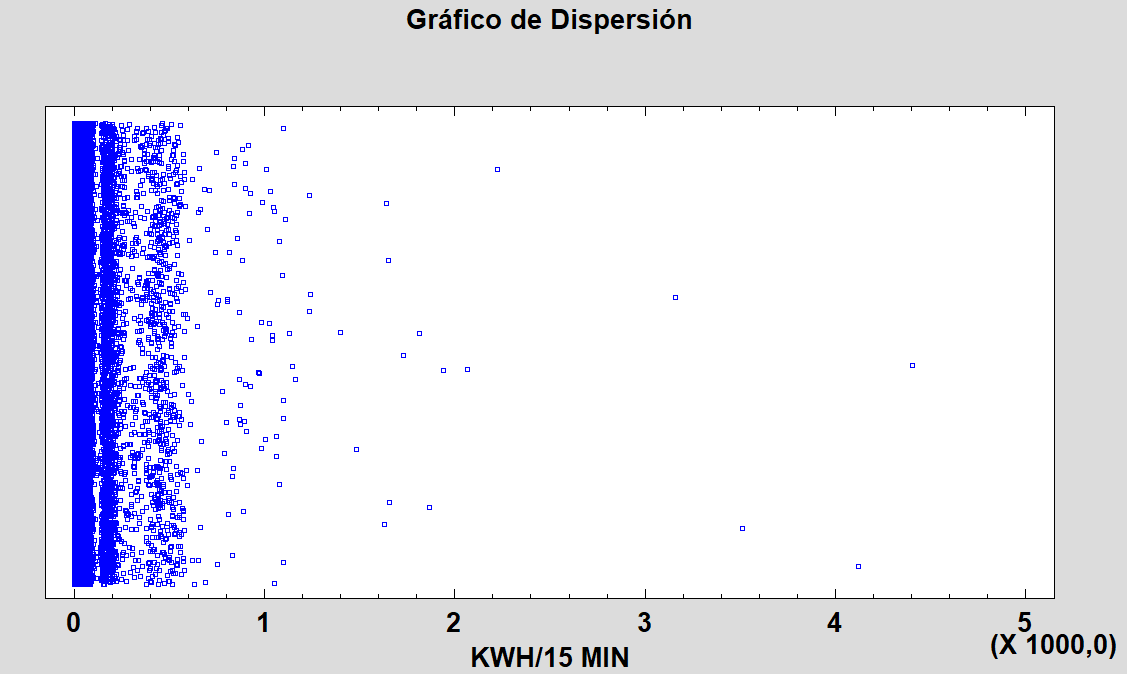
Datos antes de la limpieza:



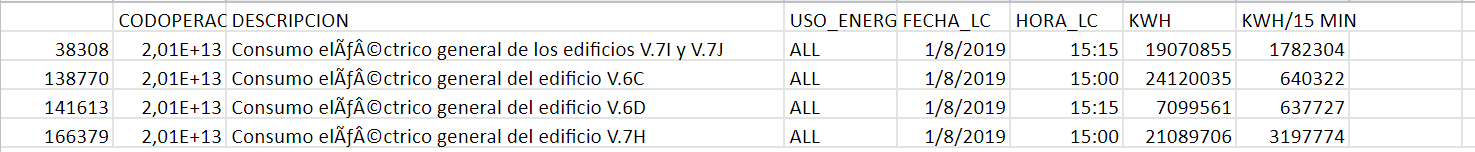


Datos después de la limpieza:





Estos han sido los datos eliminados en Agosto:



Como podemos ver, la media y la desviación típica se han reducido en 30 y 8580 puntos, mientras que los parámetros robustos se han mantenido igual. Los valores eliminados son tan altos que eclipsan a todos los demás en el gráfico de dispersión y de aberrantes.

Como curiosidad, los datos atípicos de este mes vienen todos de la zona “V”, alrededor del mismo día y a la misma hora