EVALUACION ESTADISTICA DE LA TASA DE PROCESAMIENTO Y DEMANDA DE POTENCIA EN MOLIENDA SAG

Emmanuel Herrera Flores

15 Noviembre del 2018

OBJETIVO

Realizar un análisis estadístico respecto a la evolución del tonelaje total procesado por los molinos SAG n°16 y 17 de la planta de molienda A-2 de la concentradora de División Chuquicamata (DCHU), con el fin de encontrar (o no) diferencias estadísticamente significativas de la variables en estudio en períodos a definir.

METODOLOGIA

Con el propósito de entender la variabilidad operacional de la sección de molienda A-2 se extrajo data de operación histórica desde Pi System en un período que consideró el 01-01-2017 hasta el 16-10-2018 (22 meses de operación) en una frecuencia de 5 minutos.

La data se extrajo en modalidad "sample data" con lo cual se evito cualquier pre procesamiento de esta.

De acuerdo a lo solicitado por el cliente, el análisis consideró los siguientes períodos:

- 1er período: 01-01-2017 hasta 31-10-2017
- 2do período: 01-11-2017 hasta 31-08-2018
- 3er período: 01-09-2018 hasta 16-10-2018

DESCRIPCION DEL PROCESO

La sección de molienda A-2 de Division Chuquicamata (DCHU) esta compuesta por dos lineas de molienda tipo SAG, cada una con dos molinos de bolas (BM) operando en circuito inverso. Los Pebbles generados por ambas líneas SAG son chancados en una planta dedicada, enviando el producto al molino de bolas unitario de la seccion n°19 (conocido también como quinto molino). Adicionalmente este molino unitario es alimentado con pulpa proveniente de los cajones que alimentan a las BHC de ambas lineas SAG. El flowsheet de la sección de moliedna A-2 se muestra a continuación:

Los principales equipos que componene la seccion de molienda A-2 son los siguientes:

- Molinos SAG n° 16 Y 17
- Molinos de bolas n°16a, 16b, 17a y 17b
- Baterias de hidrociclones n°16a, 16b, 17a y 17b
- Cajones de alimentación a la batería de hidrociclones
- Planta de Chancado de Pebbles, compuesta por 3 chancadores de cono
- Quinto Molino

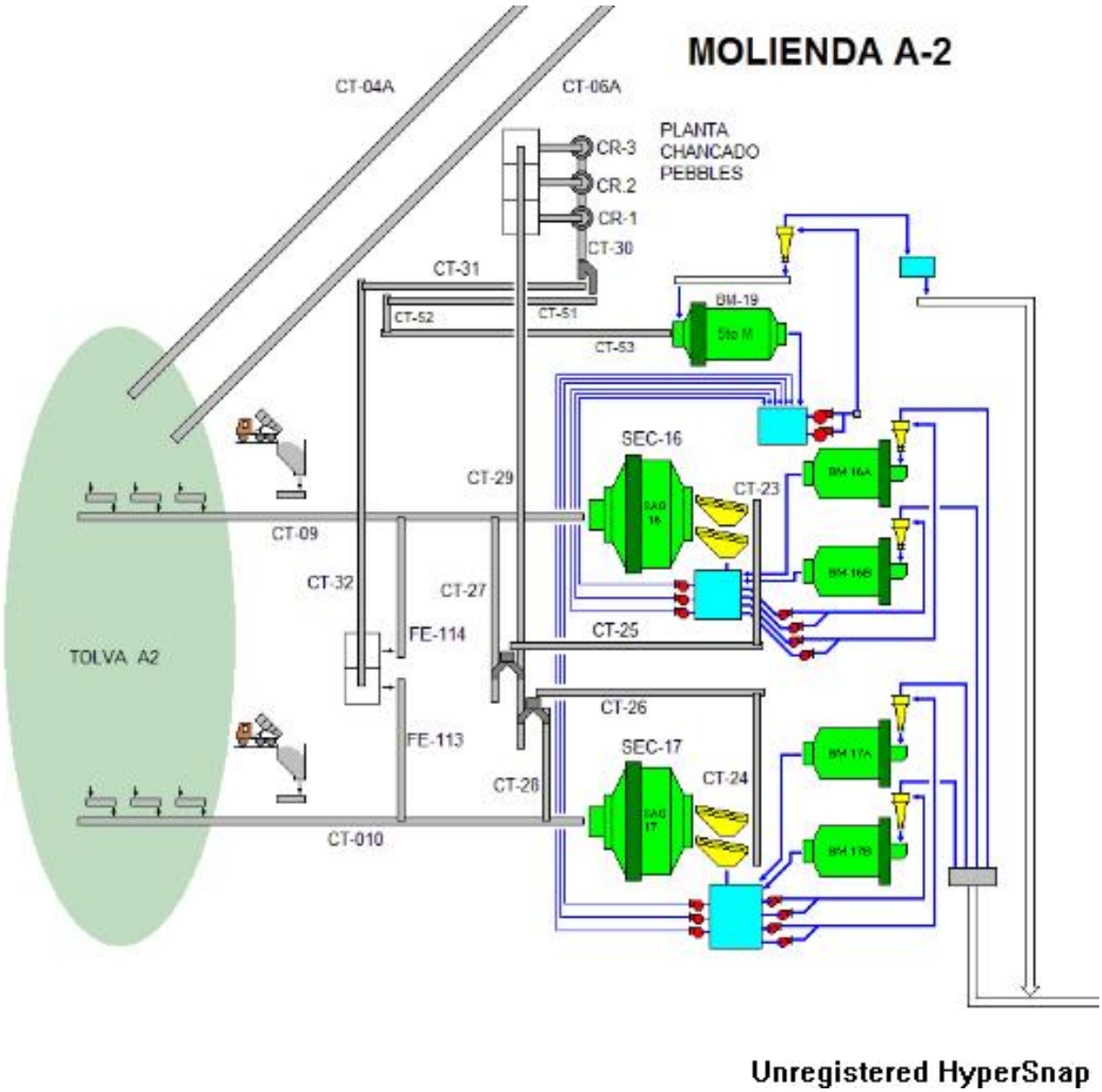

MOLIENDA A-2
CT-04A
CT-06A
PLANTA
CHANCADO
PEBBLES
CR-3
CR.2
CR-1
CT-30
CT-31
CT-52
CT-51
CT-53
BM-19
Sto M
SEC-16
CT-29
CT-23
CT-09
CT-32
CT-27
FE-114
CT-25
TOLVA A2
CT-26
FE-113
SEC-17
CT-24
CT-28
CT-010
Unregistered HyperSnap

Figure 1:

EXPLORACION INICIAL

Se recolectaron cinco variables relativas a la solictud realizada, con un total de 188221 observaciones/datos por variable (equivalentes a 22 meses de operacion cada 5 minutos). Estas vaiables son:

- tph_f_sag: toneladas horarias frescas totales procesadas por cada molinos SAG según corresponda
- mw_sag: potencia en Mega Watts reportada por cada molino SAG según corresponda

A continuación se puede apreciar un resumen estadístico de cada una de las variables antes señaladas:

```
data_sag %>%
  select(tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  summary(digits = 3)
```

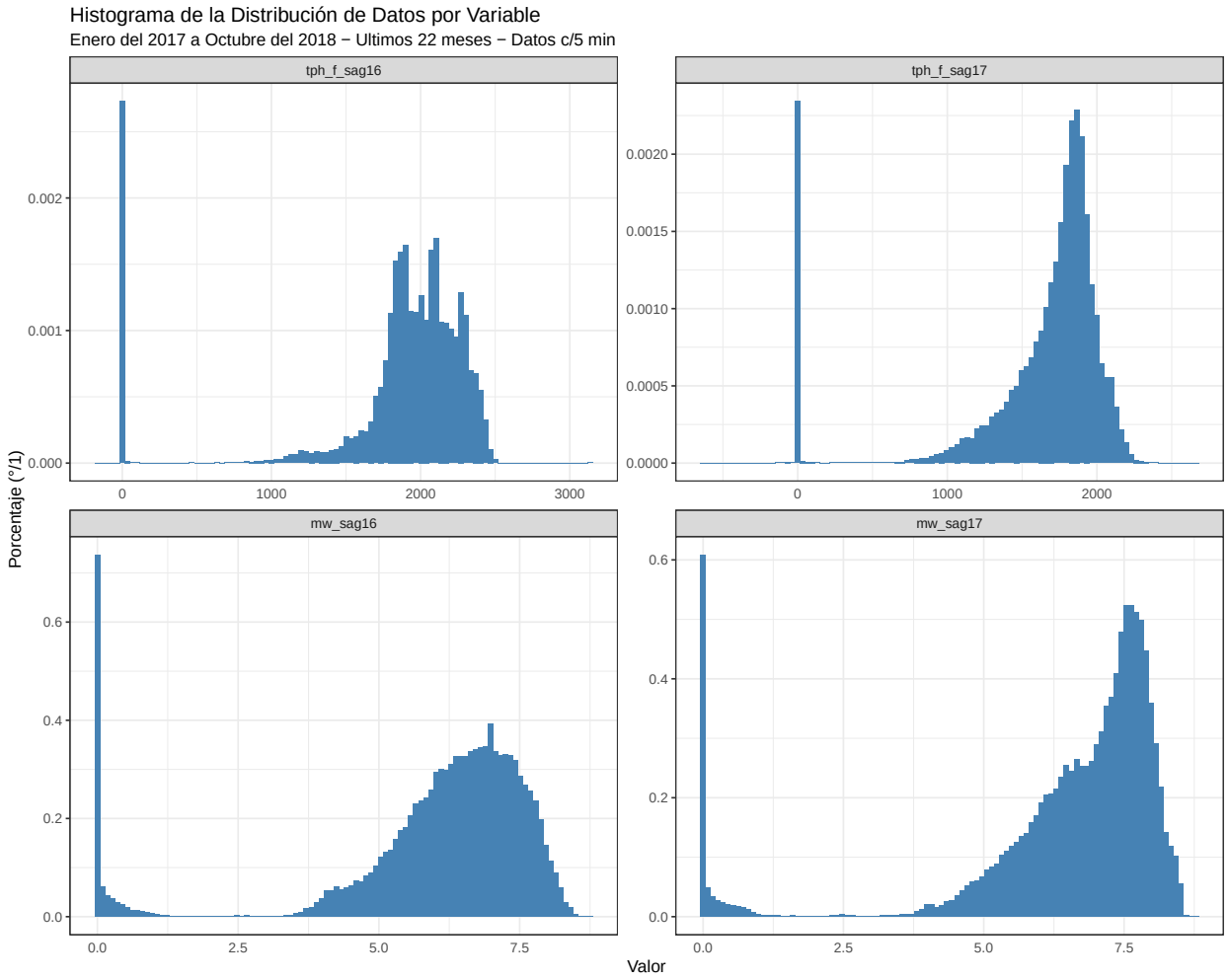
```
  tph_f_sag16    tph_f_sag17      mw_sag16       mw_sag17
 Min.   :-154   Min.   :-645   Min.   :0.00   Min.   :0.00
 1st Qu.:1794   1st Qu.:1530   1st Qu.:5.52   1st Qu.:6.05
 Median :1973   Median :1777   Median :6.44   Median :7.08
 Mean   :1807   Mean   :1598   Mean   :5.89   Mean   :6.42
 3rd Qu.:2164   3rd Qu.:1894   3rd Qu.:7.17   3rd Qu.:7.66
 Max.   :3150   Max.   :2650   Max.   :8.75   Max.   :8.78
 NA's   :2451   NA's   :2447   NA's   :2037   NA's   :1910
```

Se puede apreciar, para cada variable, los siguientes estadísticos:

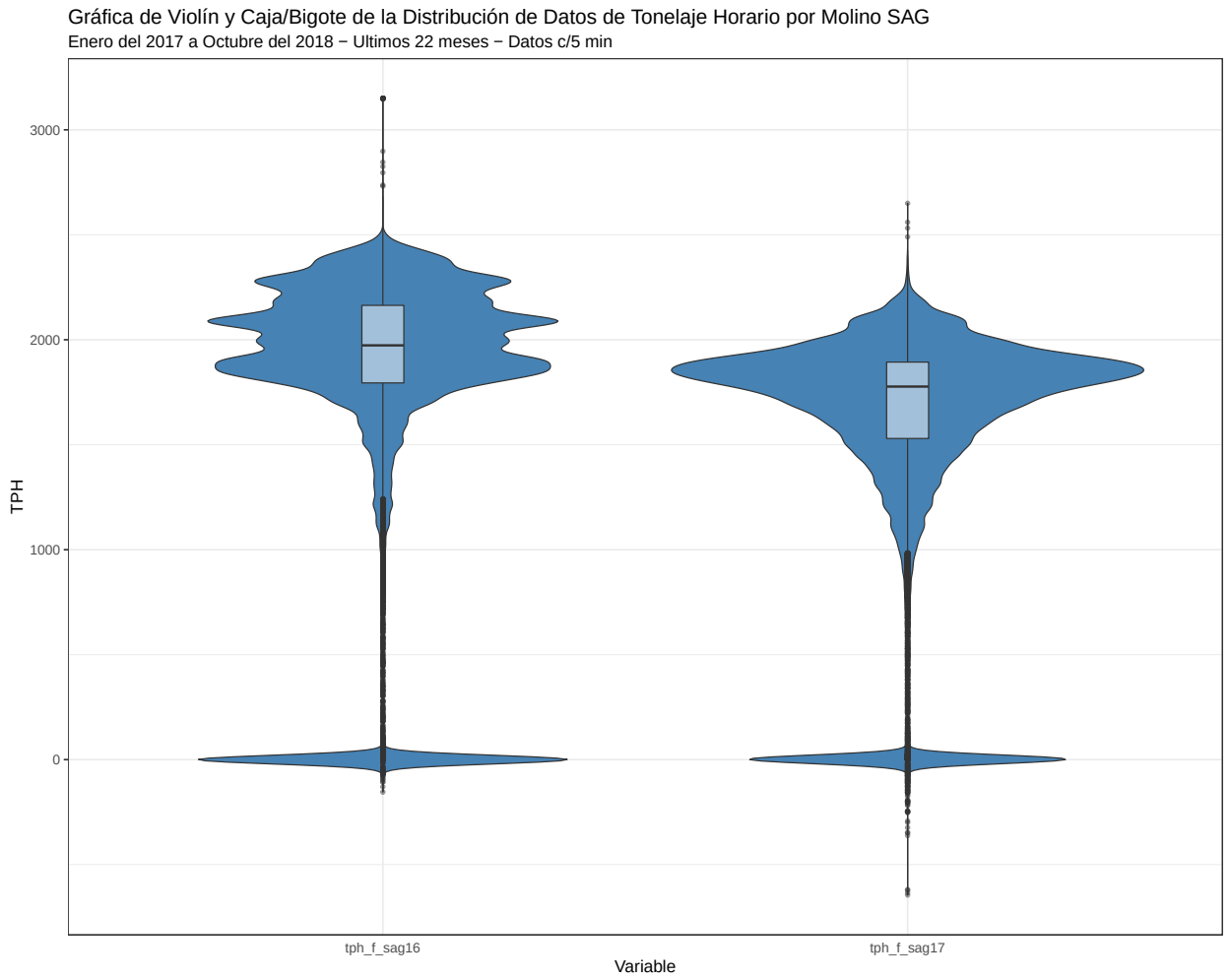
- Min: valor mínimo
- 1st Qu: primer cuartil, valor que contiene el 25% de los datos de la variable
- Median: mediana, , valor que contiene el 50% de los datos de la variable
- Mean: media o promedio
- 3rd Qu: tercer cuartil, valor que contiene el 75% de los datos de la variable
- MAx: valor máximo
- NA´s: datos perdidos. Los cuales corresponden a observaciones que debieron ser registrados pero que por diversas razones (usualmente fallas en los instrumentos de medición) no fueron leídas.

El resumen antes indicado se pude apreciar de forma gráfica a continuación:

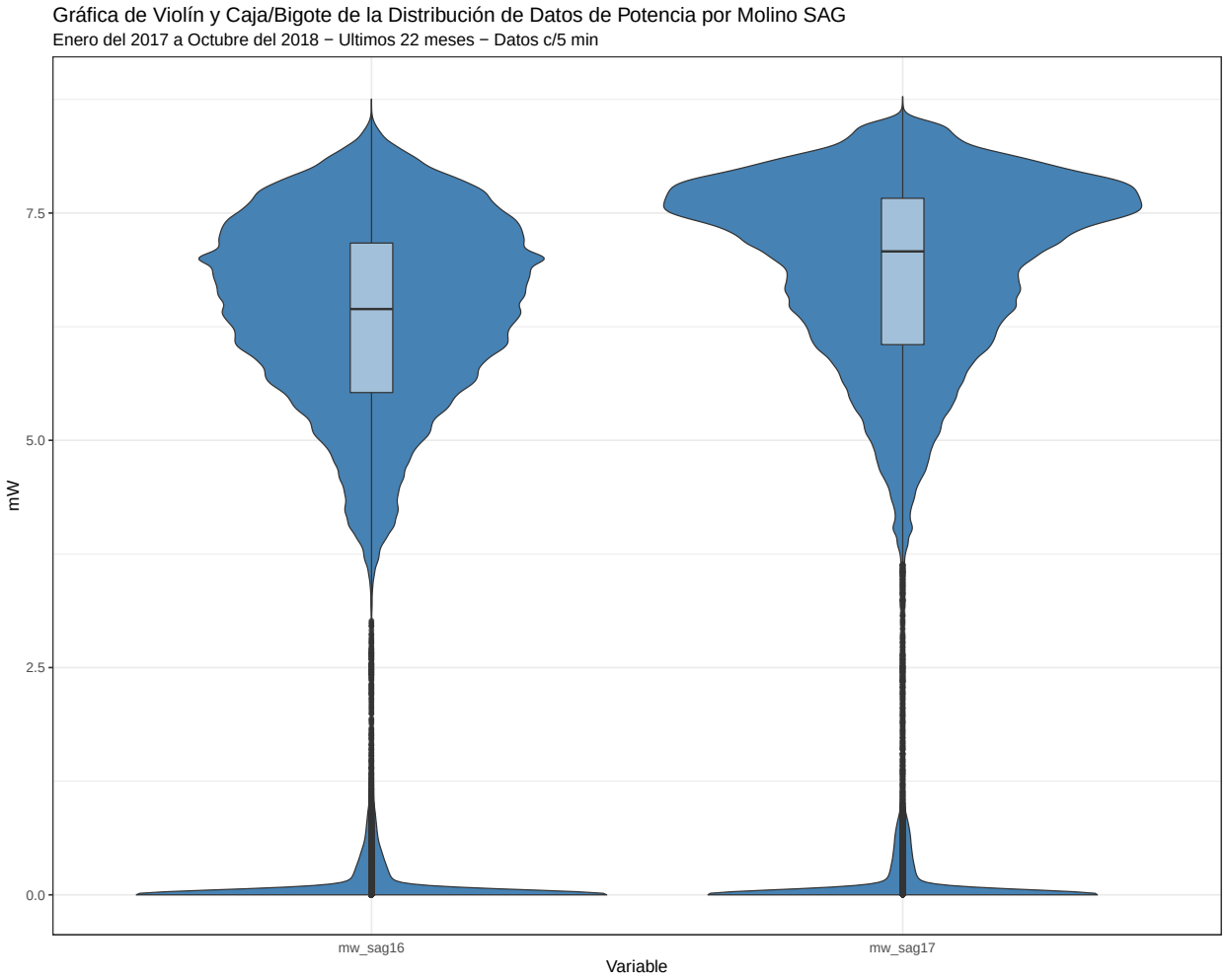
```
#Histogramas
data_sag %>%
  select(fecha, tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  na.omit() %>%
  stack() %>%
  ggplot(aes(x=values, y=stat(density), fill=values))+
  geom_histogram(bins = 100, fill='steelblue')+
  facet_wrap(~ind, scales = "free")+
  labs(title = "Histograma de la Distribución de Datos por Variable",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses - Datos c/5 min",
       y = "Porcentaje (°/1)",
       x = "Valor")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())
```




```
#Violin + boxplot: TPH
data_sag %>%
  select(tph_f_sag16, tph_f_sag17) %>%
  na.omit() %>%
  stack() %>%
  ggplot(aes(x=ind, y=values)) +
  geom_violin(bw=20, fill='steelblue')+
  geom_boxplot(alpha=0.5, width=0.08)+
   labs(title = "Gráfica de Violín y Caja/Bigote de la Distribución de Datos de Tonelaje Horario por Mo
        subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses - Datos c/5 min",
        y = "TPH",
        x = "Variable")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())
```




```
#Violin + boxplot: Potencia - mW
data_sag %>%
  select(mw_sag16, mw_sag17) %>%
  na.omit() %>%
  stack() %>%
  ggplot(aes(x=ind, y=values)) +
  geom_violin(bw=0.05, fill='steelblue')+
  geom_boxplot(alpha=0.5, width=0.08)+
  labs(title = "Gráfica de Violín y Caja/Bigote de la Distribución de Datos de Potencia por Molino SAG"
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses - Datos c/5 min",
       y = "mW",
       x = "Variable")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())
```



De los diferentes gráficos la principal característica observada radica en la existencia de una fuerte asimetría negativa (cola izquierda o inferior) de los datos, lo cual nos indica la existencia de períodos de operación cuyas mediciones se reportan hacia el intervalo izquierdo de los indicadores de tendencia central (media y mediana). Este tipo de variabilidad en la operación de los molinos SAG puede tener diferentes causas, por mencionar algunas: motivos de índole operacional que impliquen un procesamiento menor al habitual, paradas programadas, paradas no programadas, entre otras.

Los gráficos de violín + caja y bigote (boxplots) sugieren una diferencia importante en las tasas de procesamiento de los molinos SAG, en donde la mediana (equivalente al 50% de los datos) del SAG n°16 supera/contiene al 75% de las observaciones del SAG n°17. Ademas la gráfica de violín del SAG n°16 muestra una mayor concentración de datos sobre las 2000 tph, no asi el SAG n°17 el cual muestra una fuerte cocnetración en torno a las 1800 tph.

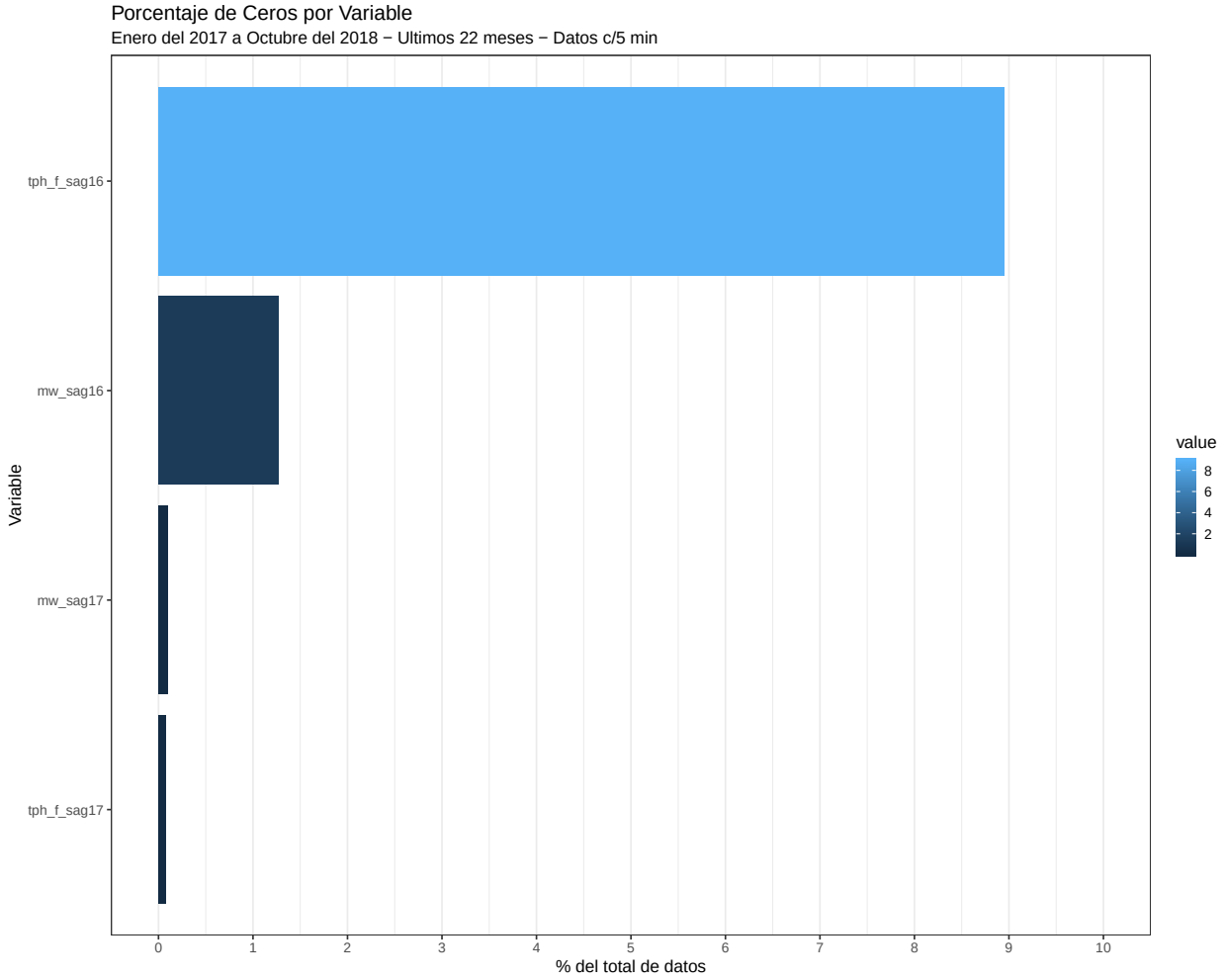
Respecto a la potencia el fenómeno se invierte, es decir, el SAG n°17 muestra un mayor consumo de potencia pero la diferencia no es tan marcada, sin embargo este molino muestra una gran concentración de mediciones en torno a los 7.5 mW. El molino SAG n°16 reporta mediciones de potencia no tan concentradas y, en general, mas dispersas.

También se observan concentraciones no despreciables de observaciones reportadas como “ceros” y/o cercanas a “cero” en todas las variables, incluso se aprecian observaciones negativas en el caso de la alimentación a los molinos SAG’s. Estas lecturas (“ceros”), pueden tener su origen ya sea en mediciones efectivas, lo que implicaría un indicativo del tiempo total que el equipo NO estubo disponible para operar (disponibilidad), o que estubo disponible para operar pero que por diversas razones no operó (utilización), o a fallas en

los instrumentos de medición en el caso de los valores negativos (en cuyo caso estas observaciones deben considerarse como NA's).

En el gráfico que se muestra a continuación se puede apreciar de forma mas clara el porcentaje del tiempo total en que las diferentes variables reportaron "ceros":

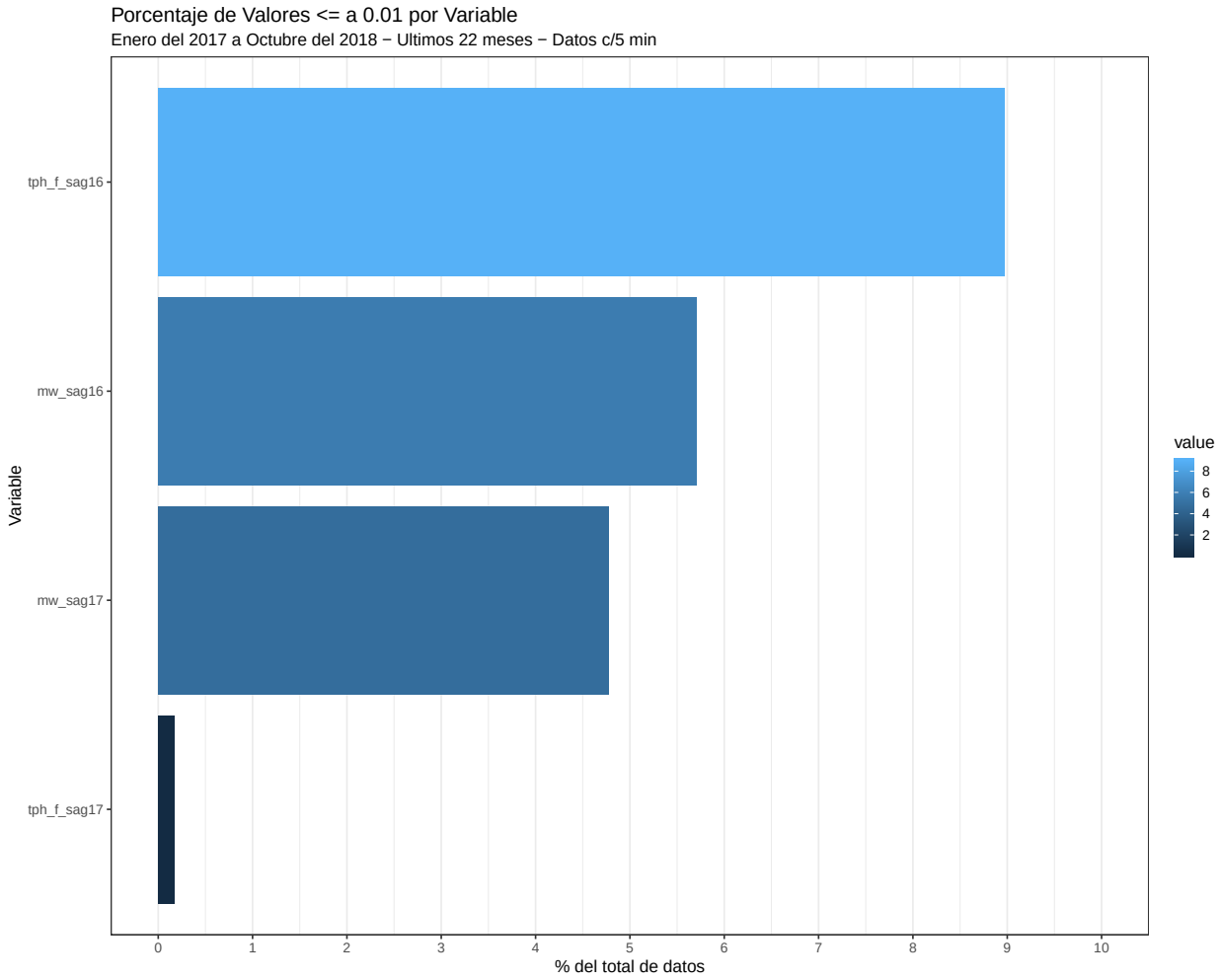
```
#Exploracion de "ceros"
data_sag %>%
  select(tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  lapply(function(x){ length(which(x==0))/length(x)*100}) %>%
  melt() %>%
  ggplot(aes(x=reorder(L1,value), y=value, fill=value))+
  geom_col()+
  coord_flip()+
  scale_y_continuous(breaks = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9,10),  limits=c(0, 10))+
  theme(panel.grid.major.y = element_blank())+
  labs(title = "Porcentaje de Ceros por Variable",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses - Datos c/5 min",
       y = "% del total de datos",
       x = "Variable")+
  theme_bw(base_size = 15)+
  theme(panel.grid.major.y = element_blank())
```



Se pude apreciar que, por ejemplo, la variable "tph_sag16" reporta aproximadamente un 8.9% de "ceros" correspondiente a 16751.669 observaciones/datos, siendo lo anterior extrapolable para el resto de las variables.

Cabe señalar que adicionalmente se reporta una cantidad no menor de observaciones cuyos valores están cerca de cero y/o son negativos, los cuales tambien influyen en el grado de asimetría negativa observado en las gráficas antes mostradas. Por ende a continuación se muestran el porcentaje del tiempo total en que las diferentes variables reportaron valores menores e iguales a "0.01":

```
#Exploracion de "ceros"
data_sag %>%
  select(tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  lapply(function(x){ length(which(x<=0.01))/length(x)*100}) %>%
  melt() %>%
  ggplot(aes(x=reorder(L1,value), y=value, fill=value))+
  geom_col()+
  coord_flip()+
  scale_y_continuous(breaks = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9,10),  limits=c(0, 10))+
  theme(panel.grid.major.y = element_blank())+
  labs(title = "Porcentaje de Valores <= a 0.01 por Variable",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses - Datos c/5 min",
       y = "% del total de datos",
       x = "Variable")+
  theme_bw(base_size = 15)+
  theme(panel.grid.major.y = element_blank())
```



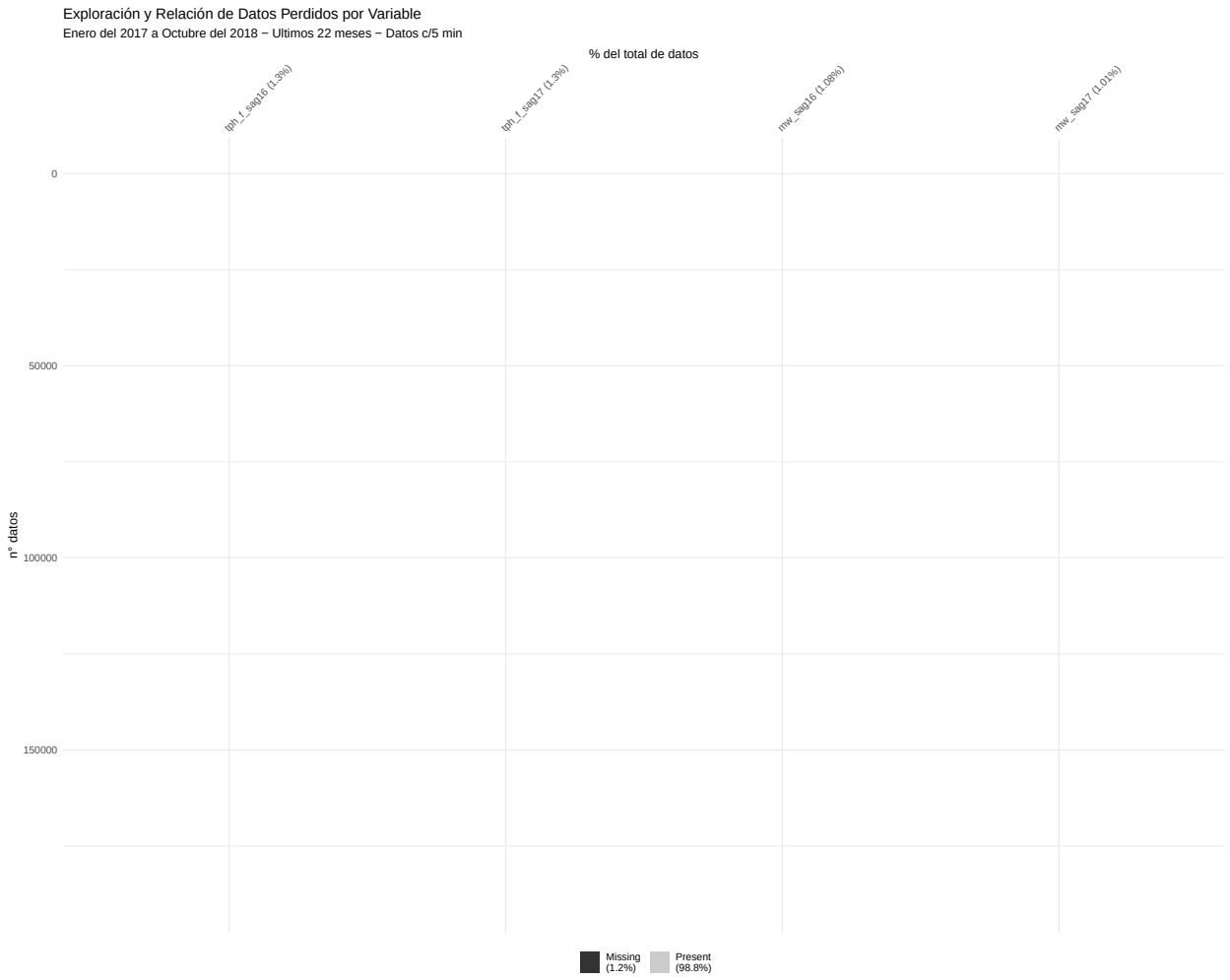
Se aprecia que el porcentaje de datos en este intervalo de valores sube de forma importante, en especial para las mediciones de potencia, las cuales reportan muchos valores cercanos a cero, los cuales podrían ser causa de mediciones en donde el equipo se encontraba en proceso de detención o marcha.

Otro punto importante a considerar tiene razón con los datos perdidos (NA's), los cuales corresponden a observaciones que debieron ser registrados pero que por diversas razones (usualmente fallas en los instrumentos de medición) no fueron leídos.

En general estos datos perdidos/NA's (en función de la cantidad que exista) se pueden reemplazar o descartar, aunque es de vital importancia considerar que los métodos de eliminación y/o reemplazo que se utilizen, pueden tener efectos determinantes sobre las conclusiones extraídas desde los análisis posteriores.

Un resumen gráfico de este tipo de observaciones se muestra a continuación:

```
#Exploracion y Relacion de datos perdidos
data_sag %>%
  select(tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  vis_miss(sort_miss=TRUE,show_perc = TRUE)+
  labs(title = "Exploración y Relación de Datos Perdidos por Variable",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses - Datos c/5 min",
       y = "n° datos",
       x = "% del total de datos")
```



Se puede apreciar que tanto para el total de los datos como para cada variable por separado el % de datos perdidos se mueve en torno al 1%, además la mayoría de los datos perdidos (líneas negras horizontales) son comunes a las 4 variables.

Como regla empírica cuando el total de datos perdidos es menor al 5% en cada variable es pausible eliminarlos sin causar una disminución relevante en la calidad del conjunto de observaciones. Lo anterior se acentúa aun mas en nuestro caso dado el hecho de que la mayoría de estos NA's son comunes a todas las variables, con lo cual solo se estarían eliminando las filas correspondientes a la información perdida.

Sin embargo en este análisis deben considerarse además aquellas observaciones que reportaron "ceros" y valores numéricos muy distantes al resto (outliers), lo cual se aborda en el capitulo siguiente.

PRE PROCESAMIENTO DE DATOS

El pre-procesamiento de datos da cuenta de diferentes acciones que apuntan a "limpiar" el conjunto de datos de aquellas observaciones perdidas y/o anómalas, las cuales no son de interés y además podrían interferir en el análisis estadístico a realizar. Este tipo de información da cuenta de:

- Outliers o valores atípicos, esto son datos cuyo valor es muy distante al resto (o al grueso) de los datos.
- Datos que son leídos como "ceros"
- Datos negativos

- Data perdida o NA (Not Available)

1. Valores Atípicos (outliers)

Los valores atípicos dan cuenta principalmente de todos aquellos que definen la marcada asimetría negativa (cola izquierda que se extiende hasta cero) de las distribuciones de las variables en estudio . Se asume por tanto que estos datos dan cuenta, en la mayoría de los casos, de condiciones operacionales atípicas que estan fuera del marco de operación estable del proceso (estado estacionario).

Dado lo anterior, para cada variable, se define un rango o límite bajo y sobre el cual cualquier dato será considerado anómalo y descartado del conjunto de datos:

- Rango inferior: 1er cuartil - 3*IQR
- Rango superior: 3er cuartil + 3*IQR

En donde:

- IQR es el rango intercuartílico de cada variable y se obtiene de la diferencia entre el 3er y 1er cuartil

Considerando este criterio, los rangos de operación válidos, por cada variable serían:

```
c<-c("tph16_inf", "tph16_sup", "tph17_inf", "tph17_sup", "mw16_inf", "mw16_sup", "mw17_inf", "mw17_sup"
f<-c(q_tph_f_sag16,q_tph_f_sag17,q_mw_sag16,q_mw_sag17)
c
```

```
[1] "tph16_inf" "tph16_sup" "tph17_inf" "tph17_sup" "mw16_inf"  "mw16_sup"
[7] "mw17_inf"  "mw17_sup"
```

```
print(f, digits=6)
```

```
       25%         75%         25%         75%         25%         75%
 686.270020 3271.947571  437.999512 2985.300537    0.588633   12.105753
       25%         75%
   1.225939   12.489693
```

Adicionalmente el resumen estadístico actualizado de las variables en estudio se muestra a continuación:

```
#Nuevo df sin outlier/cers/NA/nagativos e identificado segun periodod de interes
data1<-data_sag %>%
  select(fecha, tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  mutate(id=c(rep("1er período", 87552), rep("2do período", (175104-87552)), rep("3er período", (188221
  filter(tph_f_sag16>=q_tph_f_sag16[1]&tph_f_sag16<=q_tph_f_sag16[2], tph_f_sag17>=q_tph_f_sag17[1]&tph
         mw_sag16>=q_mw_sag16[1]&mw_sag16<=q_mw_sag16[2], mw_sag17>=q_mw_sag17[1]&mw_sag17<=q_mw_sag17[

summary(data1[,2:5])
```

```
  tph_f_sag16    tph_f_sag17      mw_sag16         mw_sag17
 Min.   : 687   Min.   : 445   Min.   :0.5955   Min.   :1.330
 1st Qu.:1848   1st Qu.:1620   1st Qu.:5.8145   1st Qu.:6.388
 Median :2010   Median :1798   Median :6.5651   Median :7.228
 Mean   :1991   Mean   :1737   Mean   :6.4564   Mean   :6.988
 3rd Qu.:2179   3rd Qu.:1900   3rd Qu.:7.2252   3rd Qu.:7.713
 Max.   :2898   Max.   :2650   Max.   :8.7543   Max.   :8.781
```

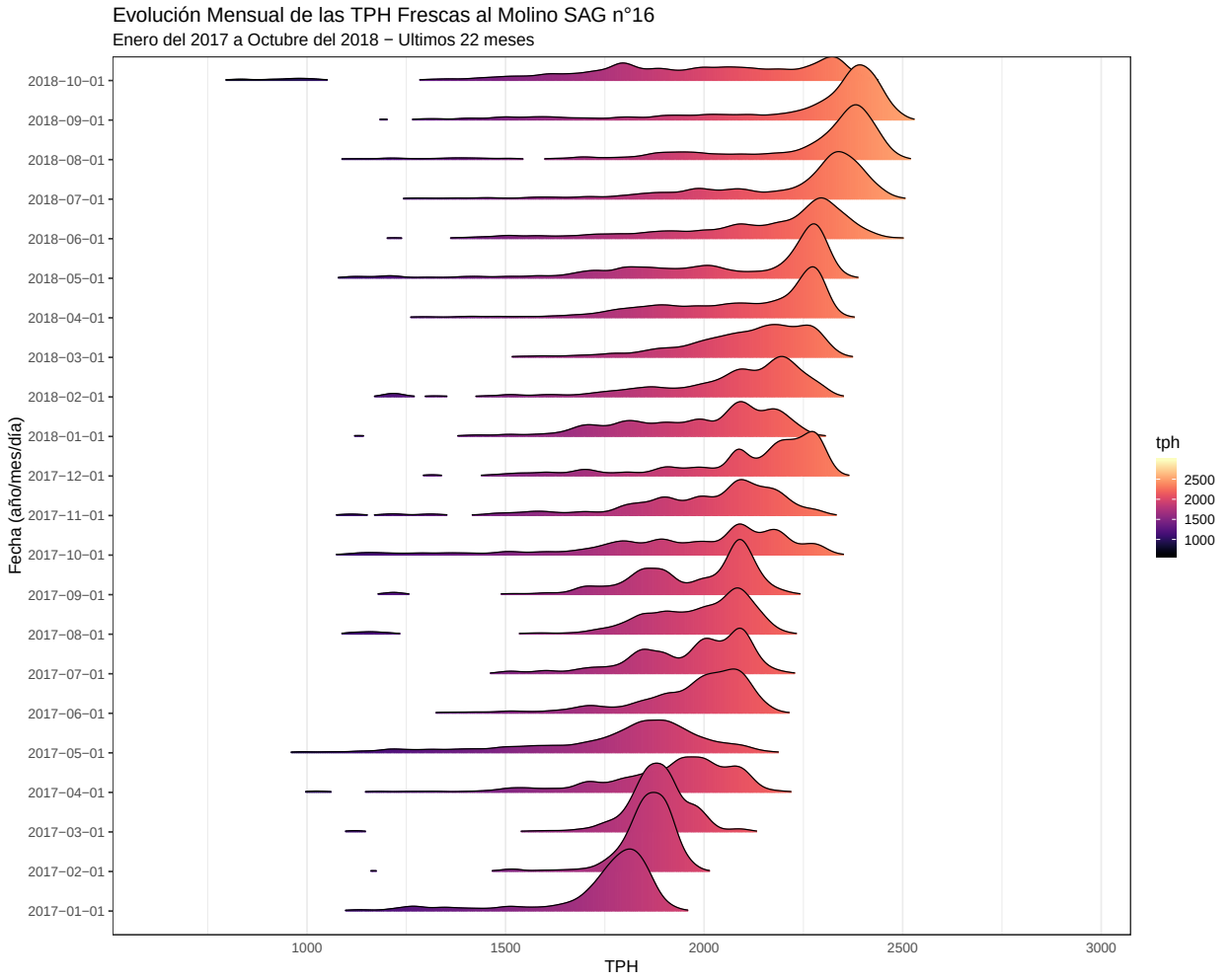
Cabe señalar que al descartar los datos extremos de la forma antes indicada se incurrió en la eliminación de todos los ceros, nagativos y NA's, por ende el tratamiento exclusivo de estos ya no es necesario.

Sin embargo al ejecutar este criterio se descartaron 32743 observaciones equivalente al 17.3960398% del conjunto de datos original.

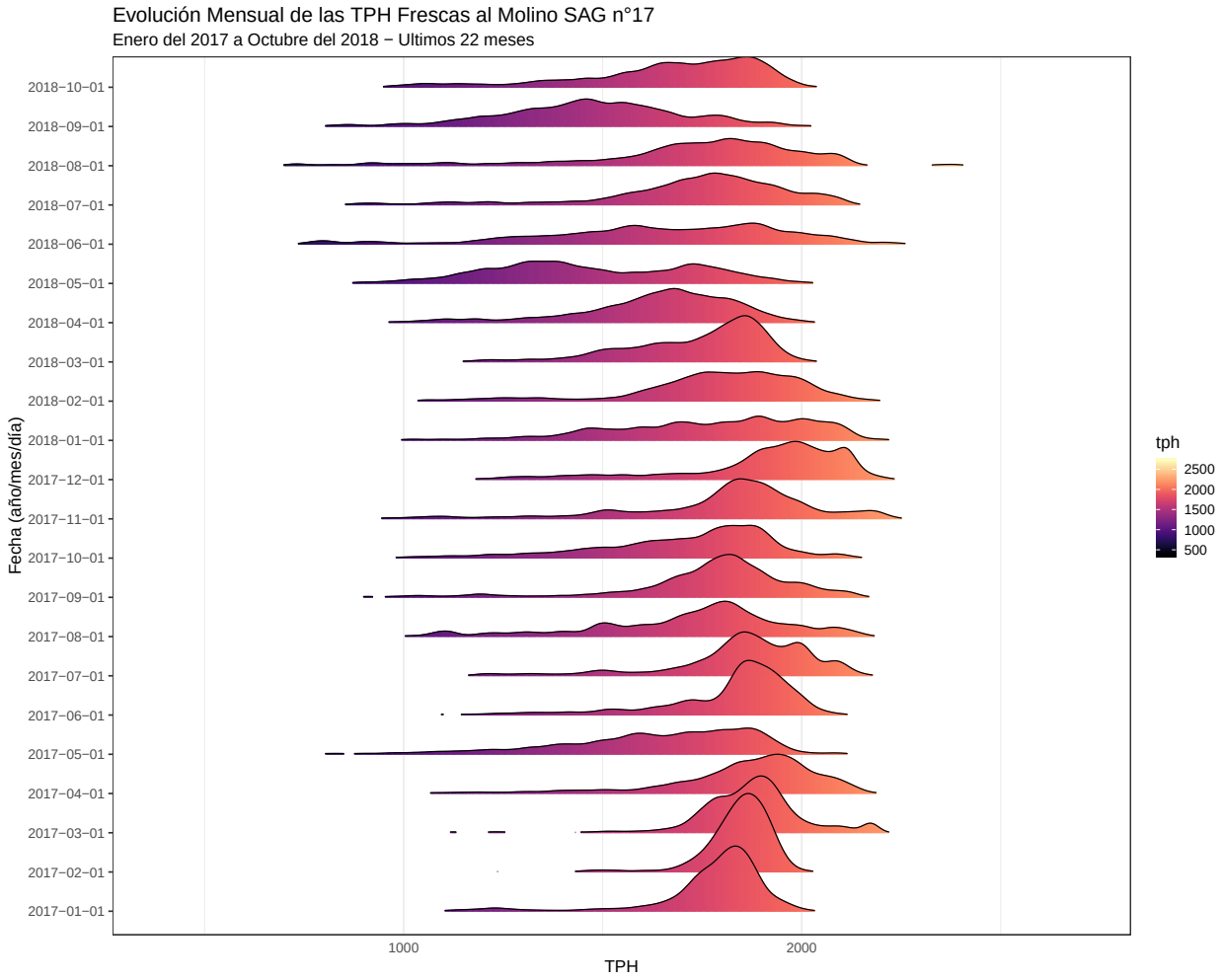
ANALISIS ESTADISTICO GRAFICO DESCRIPTIVO

Con la información pre-procesada y lista para análisis se procedió a visualizar las evoluciones mensuales de los datos para el tratamiento horario y la potencia, cuyos gráficos se muestran a continuación:

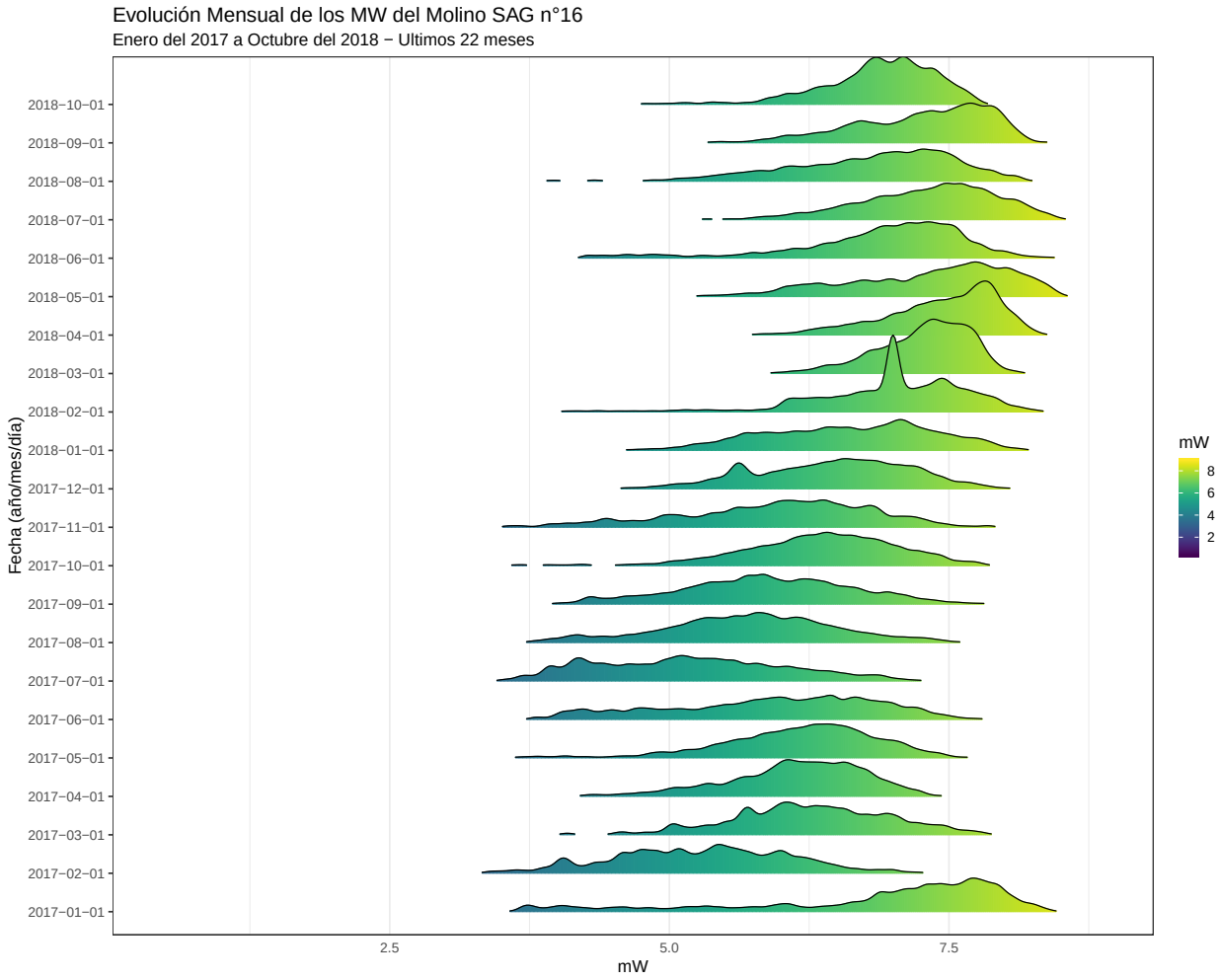
```
# Grafico temporal: evolucion promedio mensual tph sag 16
data1 %>%
  select(fecha, tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  na.omit() %>%
  group_by(fecha=floor_date(fecha, "month"))  %>%
  mutate(fecha_f=as.character(fecha)) %>%
  ggplot(aes(x= tph_f_sag16, y=fecha_f, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=20)+
  scale_fill_viridis(name = "tph", option = "A") +
  labs(title = "Evolución Mensual de las TPH Frescas al Molino SAG n°16",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses",
       y = "Fecha (año/mes/día)",
       x = "TPH")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```



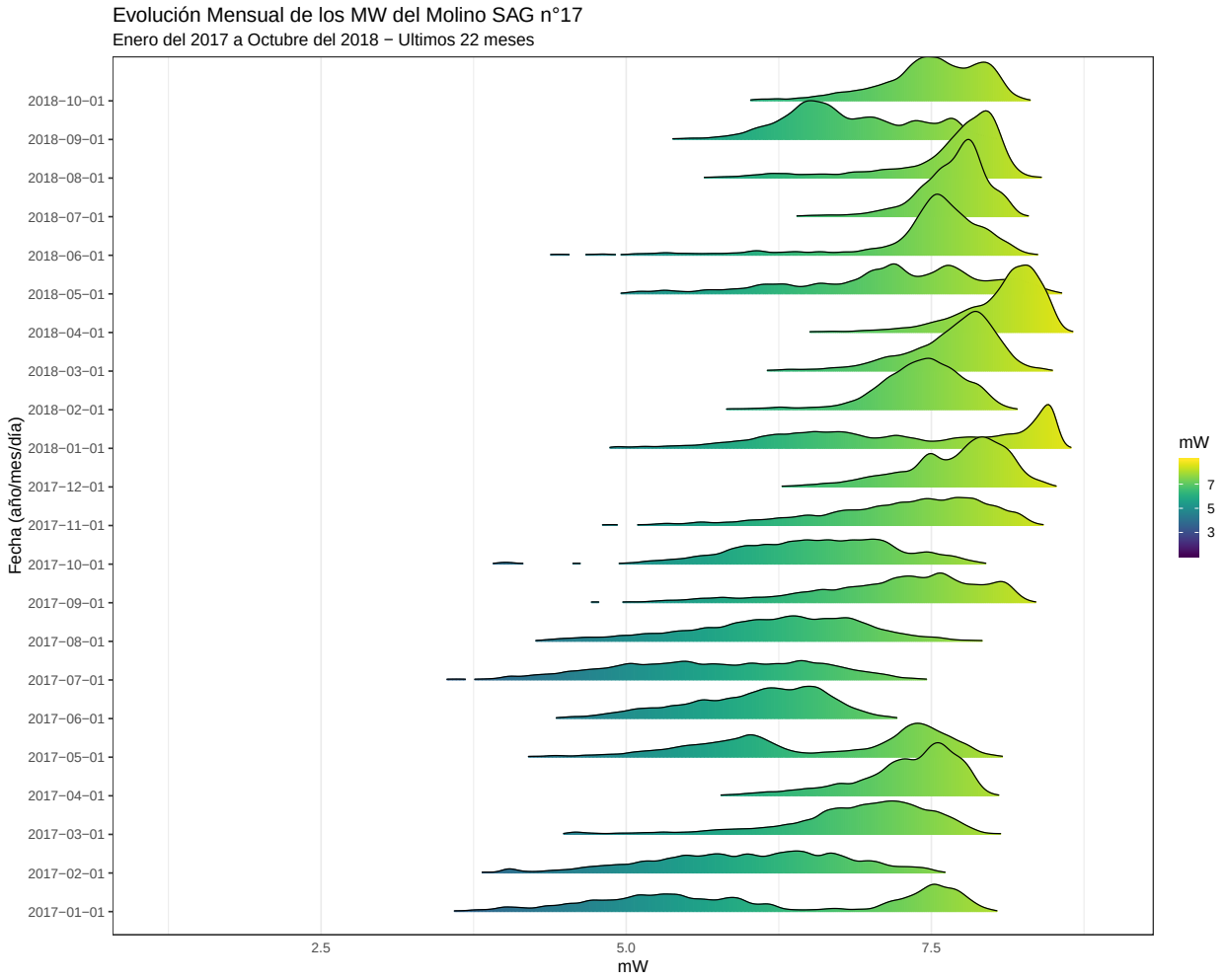

```
# Grafico temporal: evolucion promedio mensual tph sag 17
data1 %>%
  select(fecha, tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  na.omit() %>%
  group_by(fecha=floor_date(fecha, "month"))  %>%
  mutate(fecha_f=as.character(fecha)) %>%
  ggplot(aes(x= tph_f_sag17, y=fecha_f, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=20)+
  scale_fill_viridis(name = "tph", option = "A") +
  labs(title = "Evolución Mensual de las TPH Frescas al Molino SAG n°17",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses",
       y = "Fecha (año/mes/día)",
       x = "TPH")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```




```
# Grafico temporal: evolucion promedio mensual potencia sag 16
data1 %>%
  select(fecha, tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  na.omit() %>%
  group_by(fecha=floor_date(fecha, "month"))  %>%
  mutate(fecha_f=as.character(fecha)) %>%
  ggplot(aes(x= mw_sag16, y=fecha_f, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=0.05)+
  scale_fill_viridis(name = "mW", option = "D") +
  labs(title = "Evolución Mensual de los MW del Molino SAG n°16",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses",
       y = "Fecha (año/mes/día)",
       x = "mW")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```




```
# Grafico temporal: evolucion promedio mensual potencia sag 17
data1 %>%
  select(fecha, tph_f_sag16, tph_f_sag17, mw_sag16, mw_sag17) %>%
  na.omit() %>%
  group_by(fecha=floor_date(fecha, "month"))  %>%
  mutate(fecha_f=as.character(fecha)) %>%
  ggplot(aes(x= mw_sag17, y=fecha_f, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=0.05)+
  scale_fill_viridis(name = "mW", option = "D") +
  labs(title = "Evolución Mensual de los MW del Molino SAG n°17",
       subtitle = "Enero del 2017 a Octubre del 2018 - Ultimos 22 meses",
       y = "Fecha (año/mes/día)",
       x = "mW")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```



Para el sAG n°16 se puede apreaciar un incremento sostenido en los TPH frecos a traves de todo el período estudiado, con valores máximos, al inicio del período, fluctuando alrrededor de las 2000 tph y llegando, al final del período, a las 2500 tph aproximadamente.

Respecto al SAG n°17 se puede observar un comportamiento relativamente constante los primeros 12 meses (01-2017 al 12-2017), seguidos de 9 meses a la baja (01-2018 al 09-2018), con una caída sostenida en los TPH procesados y además con un aumento en la variabilidad de las mediciones (dispersión de la distribución).

En referencia a la potencia consumida por el SAG n°16 se puede observar que el primer mes (01-2017) reporta un alto consumo y dispersión, seguido de un período de inestabilidad de bajo consumo que dura aproximadamente 12 meses (hasta el 12-2017), retomando el consumo mostrado en el primer mes, el resto del período.

El molino SAG n°17 reporta consumos de potencia similares en todo el lapso de tiempo medido (salvo del 06 al 08 del 2017), en donde una principal característica da cuenta de una alta asimetría negativa.

Para todos los gráficos es importante señalar que el último mes solo considera 16 días.

ANALISIS POR PERIODO

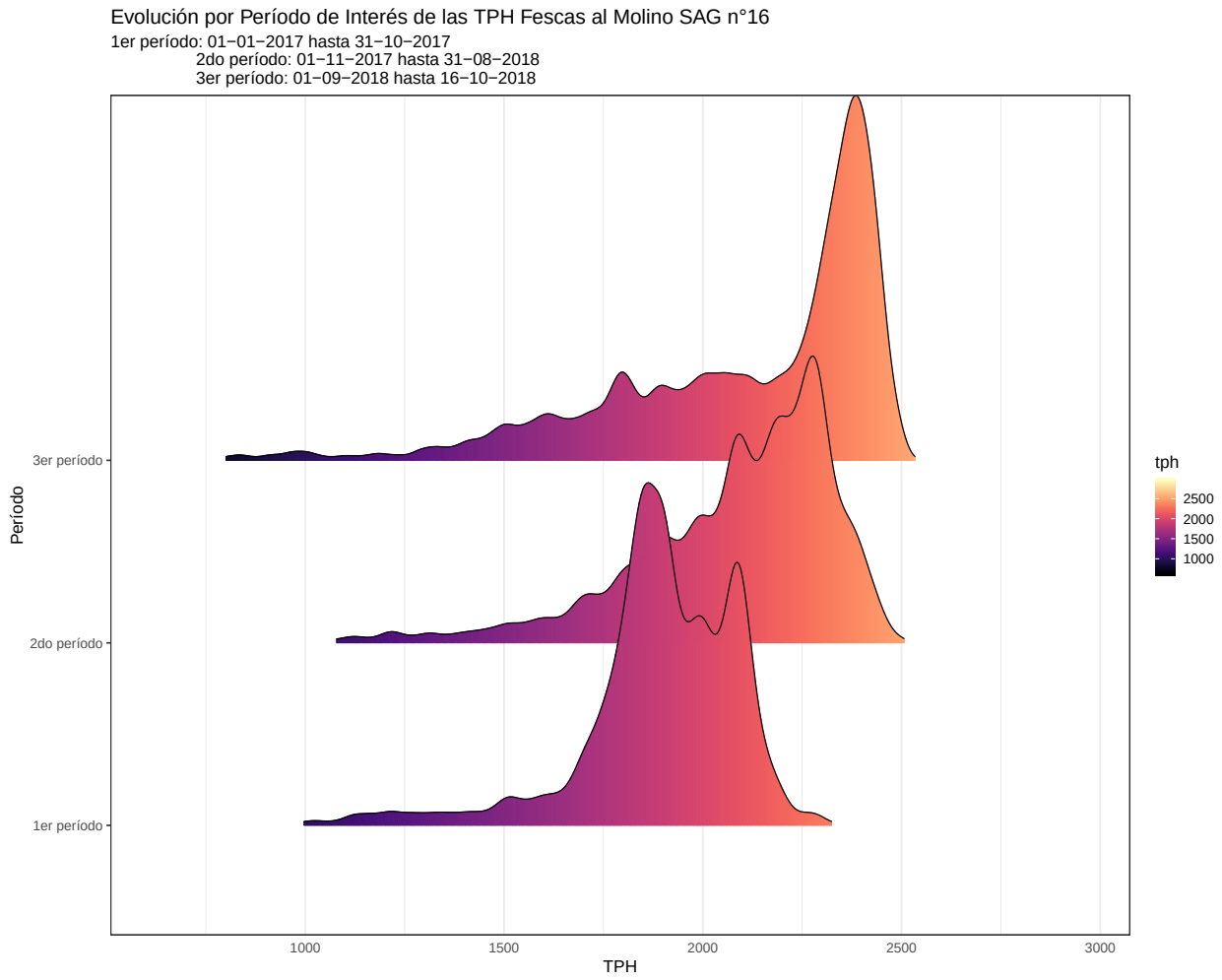
De acuerdo a lo solicitado por el cliente, el análisis consideró los siguientes períodos:

- 1er período: 01-01-2017 hasta 31-10-2017

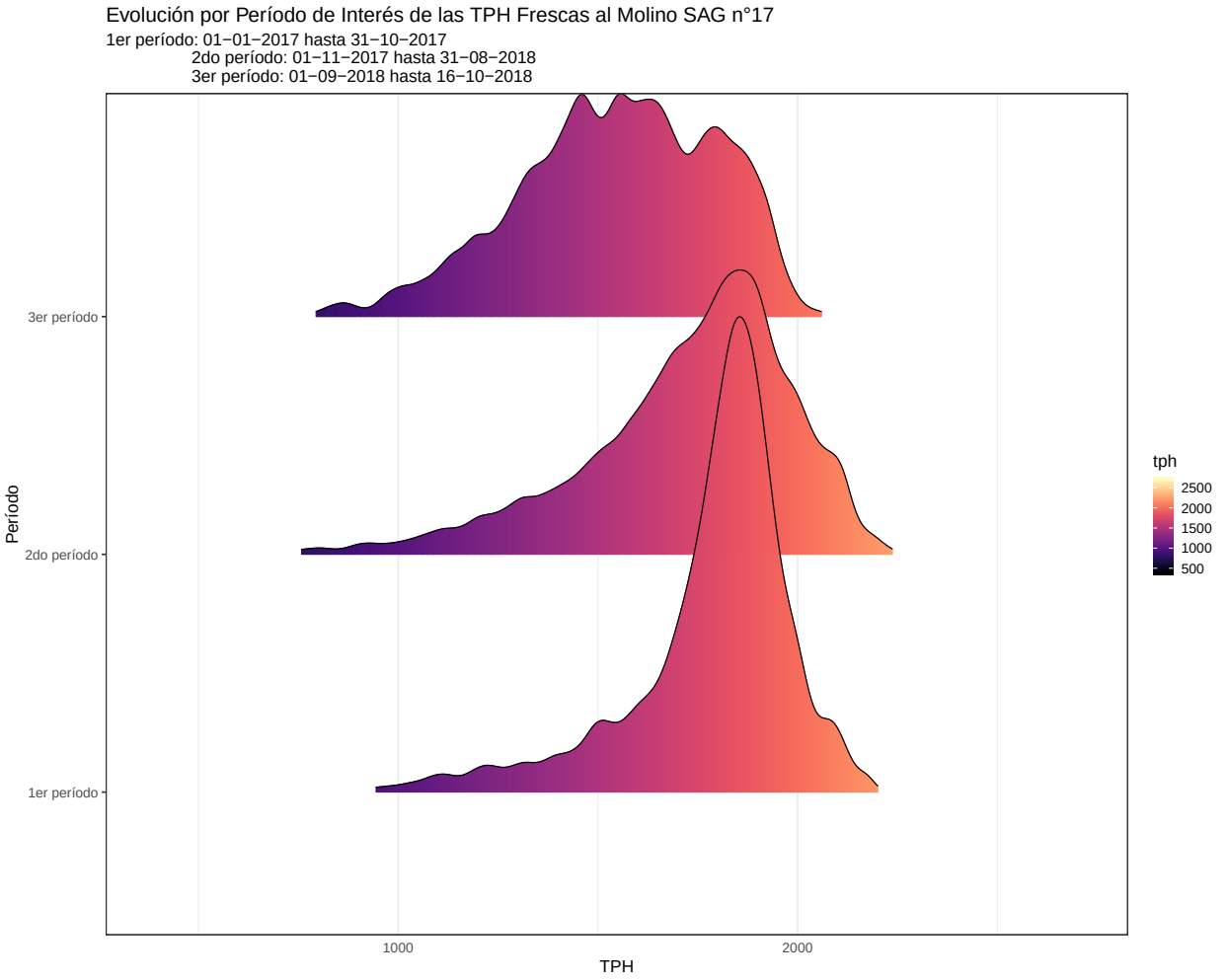
- 2do período: 01-11-2017 hasta 31-08-2018
- 3er período: 01-09-2018 hasta 16-10-2018

Por lo tanto las gráficas antes discutidas se modificaron con el fin de capturar estos intervalos de tiempo. Las gráficas se muestran a continuación:

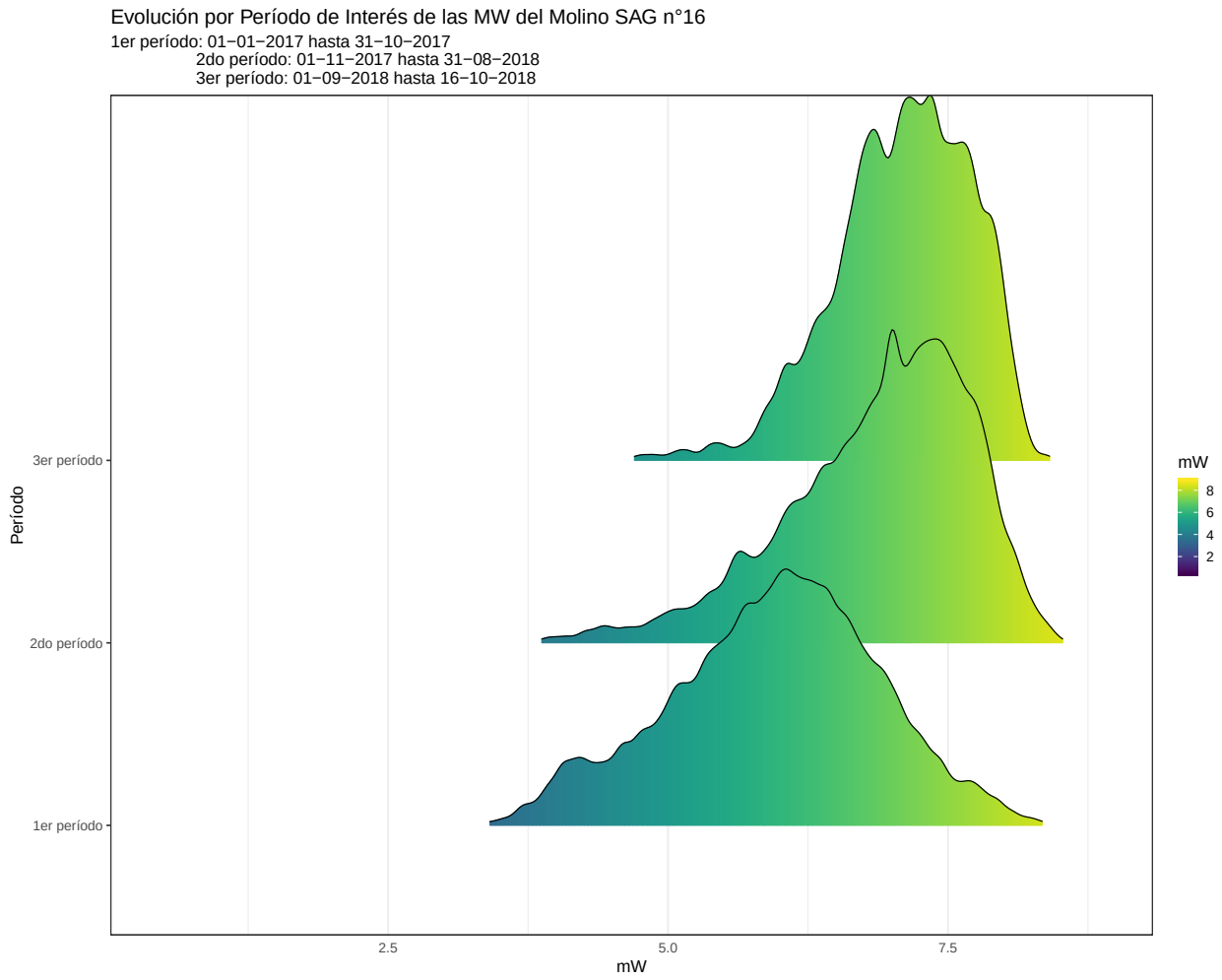
```
#Sin control de Potencia 01-01-2017 hasta 31-10-2017
#Con control de Potencia 01-11-2017 hasta 31-08-2018
#Sin control de Potencia 01-09-2018 hasta 16-10-2018

#data2 <- data1 %>%
#  mutate(id=c(rep("1er período", 72150), rep("2do período", 73604), rep("3er período", 10295)))

#TPH sag 16 por periodo
data1 %>%
  ggplot(aes(x= tph_f_sag16, y=id, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=20)+
  scale_fill_viridis(name = "tph", option = "A") +
  labs(title = "Evolución por Período de Interés de las TPH Fescas al Molino SAG n°16",
                  subtitle ="1er período: 01-01-2017 hasta 31-10-2017
                  2do período: 01-11-2017 hasta 31-08-2018
                  3er período: 01-09-2018 hasta 16-10-2018",
       y = "Período",
       x = "TPH")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```



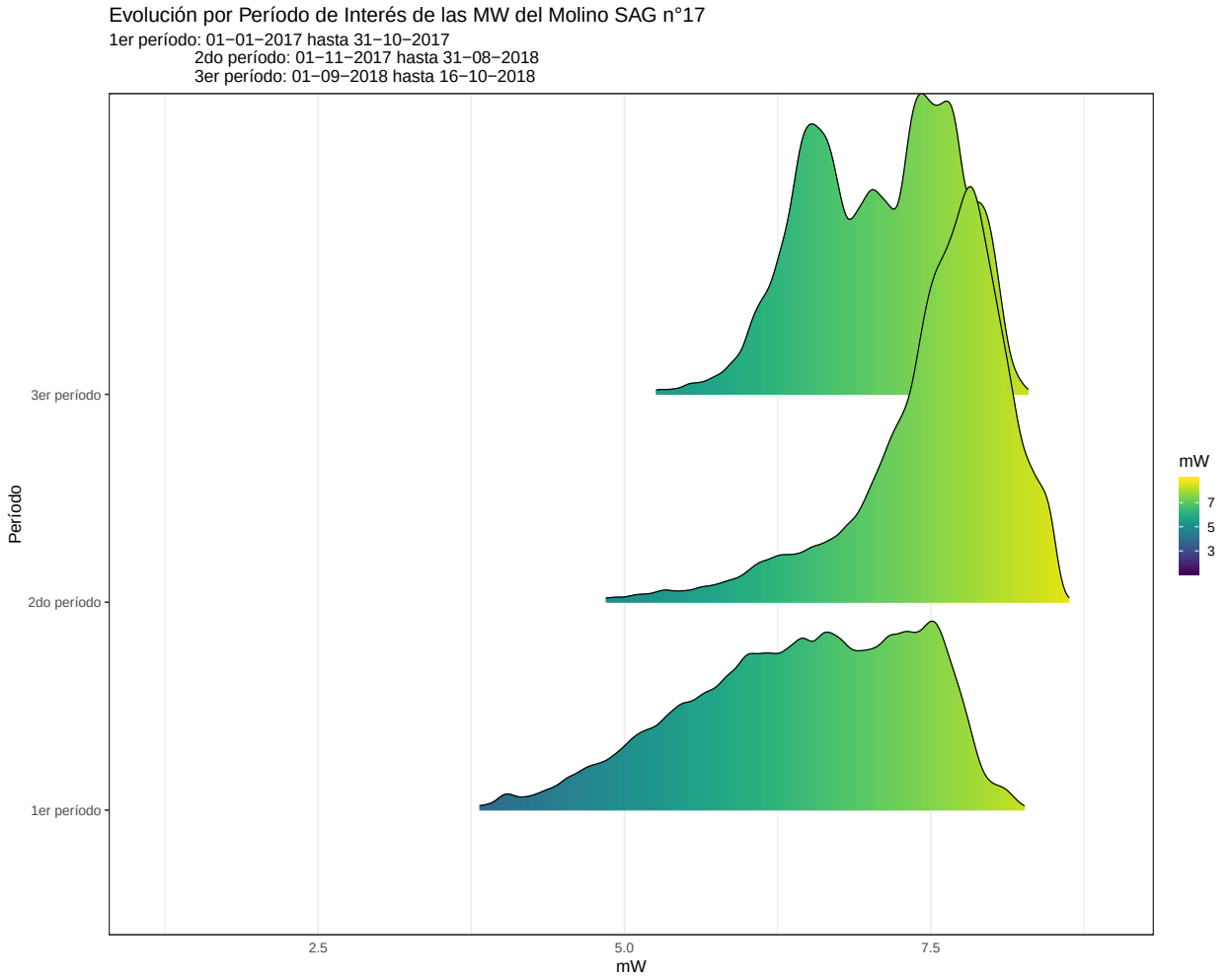

```
#TPH sag 17 por periodo
data1 %>%
  ggplot(aes(x= tph_f_sag17, y=id, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=20)+
  scale_fill_viridis(name = "tph", option = "A") +
  labs(title = "Evolución por Período de Interés de las TPH Frescas al Molino SAG n°17",
                  subtitle ="1er período: 01-01-2017 hasta 31-10-2017
                  2do período: 01-11-2017 hasta 31-08-2018
                  3er período: 01-09-2018 hasta 16-10-2018",
       y = "Período",
       x = "TPH")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```




```
#MW sag 16 por periodo
data1 %>%
  ggplot(aes(x= mw_sag16, y=id, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=0.05)+
  scale_fill_viridis(name = "mW", option = "D") +
  labs(title = "Evolución por Período de Interés de las MW del Molino SAG n°16",
                  subtitle ="1er período: 01-01-2017 hasta 31-10-2017
                  2do período: 01-11-2017 hasta 31-08-2018
                  3er período: 01-09-2018 hasta 16-10-2018",
       y = "Período",
       x = "mW")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```




```
#MW sag 17 por periodo
data1 %>%
  ggplot(aes(x= mw_sag17, y=id, fill=..x..)) +
  geom_density_ridges_gradient(scale = 2, rel_min_height = 0.01, bandwidth=0.05)+
  scale_fill_viridis(name = "mW", option = "D") +
  labs(title = "Evolución por Período de Interés de las MW del Molino SAG n°17",
                  subtitle ="1er período: 01-01-2017 hasta 31-10-2017
                  2do período: 01-11-2017 hasta 31-08-2018
                  3er período: 01-09-2018 hasta 16-10-2018",
       y = "Período",
       x = "mW")+
  theme_bw(base_size = 15)+
  theme(axis.ticks.x = element_blank())+
  theme(panel.grid.major.y = element_blank())
```



Para el sAG n°16 se puede apreciar un incremento en la concentración de las observaciones en el intervalo alto de las distribuciones, lo cual da cuenta de una asimetría negativa. También se puede apreciar que la curva del segundo período presenta un leve desplazamiento a la derecha (mayores tph) respecto principalmente del primer período.

Para el molino SAG n°17 se aprecia que el primer período muestra una alta concentración de datos en el intervalo alto de la distribución, además se visualiza que tanto para el primer y el segundo período existe un desplazamiento hacia la derecha (mayores tph).

Respecto a las mediciones de potencia se aprecia para ambos molinos un leve incremento a medida que nos movemos del primer al tercer período, este aumento pareciera amplificarse más en el segundo período, particularmente para el molino sag n° 17.

Para poder tener un mayor grado de confianza respecto a las potenciales diferencias existentes en el análisis gráfico descriptivo, a continuación se presentan diversas pruebas estadísticas cuyo propósito es confirmar (o no) la existencia diferencias estadísticamente significativas en las medias de los períodos considerados.

ANALISIS ESTADISTICO INFERENCIAL

Para corroborar la existencia de diferencias estadísticamente significativas en las medias de los períodos considerados se ocupará la herramienta estadística ANOVA (Análisis de Varianza). Las razones para ocupar

esta técnica paramétrica descansan, en primer lugar, en el hecho de que se desea realizar inferencia para las medias de tres períodos (lo cual imposibilitaría usar una prueba de tipo T Student), y en segundo lugar, porque el tamaño de la muestra permite asumir el teorema del límite central, en donde el supuesto de normalidad de los datos no es fundamental, siempre y cuando la muestra sea lo suficientemente grande (en nuestro caso es lo suficientemente grande con aprox. 150.000 datos).

El objetivo del presente análisis es probar, para cada variable en estudio, que las diferencias reportadas en los promedios para cada período son (o no) estadísticamente significativas.

ANOVA TONELAJE PROCESADO SAG n°16

El número de observaciones y medias de los tph para cada período se muestran a continuación:

```
data1 %>%
  group_by(id) %>%
  summarise(obs=n(), promedio=mean(tph_f_sag16)) %>%
  print.data.frame()
```

```
           id   obs promedio
1 1er período 72021 1886.606
2 2do período 73142 2077.151
3 3er período 10315 2111.195
```

Adicionalmente, las hipótesis contrastadas son:

- Hipótesis nula (Ho): las medias del tonelaje horario procesado es el mismo para los 3 períodos, versus
- Hipótesis alternativa (Ha): las medias del tonelaje horario procesado difieren en al menos un par de períodos.

```
##Anova 1 factor
anova_tph16<-aov(tph_f_sag16~factor(id), data = data1) #str(data2) id no es factor sino string -> se de
summary(anova_tph16)
```

```
                Df     Sum Sq   Mean Sq F value              Pr(>F)
factor(id)       2 1476776806 738388403   12300 <0.0000000000000002 ***
Residuals   155475 9333647720     60033
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En base al valor-p reportado se deduce que hay diferencias muy significativas entre los 3 períodos, por ende se puede concluir, con un 99% de confianza, que al menos existe una media que es diferente al resto.

Ahora bien el análisis de anova solo nos dice que existen diferencias significativas en las medias pero no nos indica a que grupos pertenecen esas diferencias. Por lo tanto, para identificar que períodos son diferentes a otros se hace necesario realizar pruebas complementarias de tipo PostHoc.

Prueba PostHoc TPH SAG n° 16

Para dilucidar que medias son diferentes de otras a continuación se muestran los resultados de la prueba de Tukey, para un 99% confianza (1% error):

```
TukeyHSD(anova_tph16, conf.level = 0.99)
```

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = tph_f_sag16 ~ factor(id), data = data1)

$`factor(id)`
                              diff       lwr       upr p adj
2do período-1er período 190.54557 186.79822 194.29292     0
3er período-1er período 224.58965 217.07446 232.10485     0
3er período-2do período  34.04408  26.53611  41.55206     0
```

Los p-adj reportados nos indican que se rechaza la hipótesis nula para cada comparación, es decir, las medias difieren en todos los períodos comparados (3 pares) por lo tanto se puede concluir con un 99% de confianza que las medias de todos los grupos son estadísticamente diferentes entre si.

Adicionalmente y según se mencionó al inicio, el orden de las medias es:

```
dif_tph16 <- data1 %>%
    group_by(id) %>%
    summarise(num=n(), promedio=mean(tph_f_sag16)) %>%
    arrange(desc(promedio)) %>%
    print.data.frame()
```

```
           id   num promedio
1 3er período 10315 2111.195
2 2do período 73142 2077.151
3 1er período 72021 1886.606
```

Sobre lo cual podemos concluir, con un 99% de confianza, que el Tercer Período es el que reporta la mayor tasa de procesamiento para el sag n°16, seguidos del segundo y primero respectivamente.

ANOVA TONELAJE PROCESADO SAG n°17

Los resultados se presentan a continuacion, siguiendo el mismo desarrollo realizado en la sección anterior:

```
data1 %>%
  group_by(id) %>%
  summarise(obs=n(), promedio=mean(tph_f_sag17)) %>%
  print.data.frame()
```

```
           id   obs promedio
1 1er período 72021 1777.761
2 2do período 73142 1723.392
3 3er período 10315 1548.324
```

Adicionalmente, las hipótesis contrastadas son:

- Hipótesis nula (Ho): las medias del tonelaje horario procesado es el mismo para los 3 períodos, versus

- Hipótesis alternativa (Ha): las medias del tonelaje horario procesado difieren en al menos un par de períodos.

```
anova_tph17<-aov(tph_f_sag17~factor(id), data = data1) #str(data2) id no es factor sino string -> se de
summary(anova_tph17)
```

```
                Df     Sum Sq   Mean Sq F value              Pr(>F)
factor(id)       2  500405690 250202845    4193 <0.0000000000000002 ***
Residuals   155475 9276547785     59666
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En base al valor-p reportado se deduce que hay diferencias muy significativas entre los 3 períodos, por ende se puede concluir, con un 99% de confianza, que al menos existe una media que es diferente al resto.

Prueba PostHoc TPH SAG n° 17

Para dilucidar que medias son diferentes de otras a continuación se muestran los resultados de la prueba de Tukey, para un 99% confianza (1% error):

```
TukeyHSD(anova_tph17, conf.level = 0.99)
```

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = tph_f_sag17 ~ factor(id), data = data1)

$`factor(id)`
                              diff        lwr        upr p adj
2do período-1er período  -54.36914  -58.10501  -50.63327     0
3er período-1er período -229.43739 -236.92956 -221.94522     0
3er período-2do período -175.06825 -182.55323 -167.58327     0
```

Los p-adj reportados nos indican que se rechaza la hipótesis nula para comparación, es decir, las medias difieren en todos los períodos comparados (3 pares) por lo tanto se puede concluir con un 99% de confianza que las medias de todos los grupos son estadísticamente diferentes entre si.

Adicionalmente y según se mencionó al inicio, el orden de las medias es:

```
dif_tph17 <- data1 %>%
    group_by(id) %>%
    summarise(obs=n(), promedio=mean(tph_f_sag17)) %>%
    arrange(desc(promedio)) %>%
    print.data.frame()
```

```
           id   obs promedio
1 1er período 72021 1777.761
2 2do período 73142 1723.392
3 3er período 10315 1548.324
```

Sobre lo cual podemos concluir, con un 99% de confianza, que el Primer Período es el que reporta la mayor tasa de procesamiento para el sag n°17, seguido del segundo y tercero respectivamente.

ANOVA POTENCIA CONSUMIDA SAG n°16

El número de observaciones y medias de los mW consumidos para cada período se muestran a continuación:

```
data1 %>%
  group_by(id) %>%
  summarise(obs=n(), promedio=mean(mw_sag16)) %>%
  print.data.frame()
```

```
           id   obs promedio
1 1er período 72021 5.941302
2 2do período 73142 6.873211
3 3er período 10315 7.096736
```

Adicionalmente, las hipótesis contrastadas son:

- Hipótesis nula (Ho): las medias del consumo de potencia son iguales para los 3 períodos, versus
- Hipótesis alternativa (Ha): las medias del consumo de potencia difieren en al menos un par de períodos.

```
##Anova 1 factor
anova_mw16<-aov(mw_sag16~factor(id), data = data1) #str(data2) id no es factor sino string -> se debe e
summary(anova_mw16)
```

```
                Df Sum Sq Mean Sq F value               Pr(>F)
factor(id)       2  36046   18023   23700 <0.0000000000000002 ***
Residuals   155475 118231       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En base al valor-p reportado se deduce que hay diferencias muy significativas entre los 3 períodos, por ende se puede concluir, con un 99% de confianza, que al menos existe una media que es diferente al resto.

Prueba PostHoc MW SAG n° 16

Para dilucidar que medias son diferentes de otras a continuación se muestran los resultados de la prueba de Tukey, para un 99% confianza (1% error):

```
TukeyHSD(anova_mw16, conf.level = 0.99)
```

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = mw_sag16 ~ factor(id), data = data1)

$`factor(id)`
                              diff       lwr       upr p adj
2do período-1er período 0.9319091 0.9185719 0.9452463     0
3er período-1er período 1.1554338 1.1286865 1.1821811     0
3er período-2do período 0.2235247 0.1968031 0.2502463     0
```

Los p-adj reportados nos indican que se rechaza la hipótesis nula para cada comparación, es decir, las medias difieren en todos los períodos comparados (3 pares) por lo tanto se puede concluir con un 99% de confianza que las medias de todos los grupos son estadísticamente diferentes entre si.

Adicionalmente y según se mencionó al inicio, el orden de las medias es:

```
data1 %>%
  group_by(id) %>%
  summarise(num=n(), promedio=mean(mw_sag16)) %>%
  arrange(desc(promedio)) %>%
  print.data.frame()
```

```
           id   num promedio
1 3er período 10315 7.096736
2 2do período 73142 6.873211
3 1er período 72021 5.941302
```

Sobre lo cual podemos concluir, con un 99% de confianza, que el tercer período es el que reporta el mayor consumo de potencia para el sag n°16, seguido del segundo y primero respectivamente.

ANOVA POTENCIA CONSUMIDA SAG n°17

Los resultados se presentan a continuación, siguiendo el mismo desarrollo realizado para el sag n°16:

```
data1 %>%
  group_by(id) %>%
  summarise(obs=n(), promedio=mean(mw_sag17)) %>%
  print.data.frame()
```

```
           id   obs promedio
1 1er período 72021 6.442754
2 2do período 73142 7.513387
3 3er período 10315 7.076151
```

Adicionalmente, las hipótesis contrastadas son:

- Hipótesis nula (Ho): las medias del consumo de potencia son iguales para los 3 períodos, versus
- Hipótesis alternativa (Ha): las medias del consumo de potencia difieren en al menos un par de períodos.

```
##Anova 1 factor
anova_mw17<-aov(mw_sag17~factor(id), data = data1) #str(data2) id no es factor sino string -> se debe e
summary(anova_mw17)
```

```
               Df Sum Sq Mean Sq F value              Pr(>F)
factor(id)      2  41681   20840   32897 <0.0000000000000002 ***
Residuals   155475  98496       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En base al valor-p reportado se deduce que hay diferencias muy significativas entre los 3 períodos, por ende se puede concluir, con un 99% de confianza, que al menos existe una media que es diferente al resto.

Prueba PostHoc MW SAG n° 17

Para dilucidar que medias son diferentes de otras a continuación se muestran los resultados de la prueba de Tukey, para un 99% confianza (1% error):

```
TukeyHSD(anova_mw17, conf.level = 0.99)
```

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = mw_sag17 ~ factor(id), data = data1)

$`factor(id)`
                              diff        lwr        upr p adj
2do período-1er período  1.0706323  1.0584590  1.0828056     0
3er período-1er período  0.6333967  0.6089836  0.6578099     0
3er período-2do período -0.4372356 -0.4616253 -0.4128459     0
```

Los p-adj reportados nos indican que se rechaza la hipótesis nula para cada comparación, es decir, las medias difieren en todos los períodos comparados (3 pares) por lo tanto se puede concluir con un 99% de confianza que las medias de todos los grupos son estadísticamente diferentes entre si.

Adicionalmente y según se mencionó al inicio, el orden de las medias es:

```
data1 %>%
  group_by(id) %>%
  summarise(num=n(), promedio=mean(mw_sag17)) %>%
  arrange(desc(promedio)) %>%
  print.data.frame()
```

```
          id   num promedio
1 2do período 73142 7.513387
2 3er período 10315 7.076151
3 1er período 72021 6.442754
```

Sobre lo cual podemos concluir, con un 99% de confianza, que el segundo período es el que reporta el mayor consumo de potencia para el sag n°17, seguido por el tercero y el primero respectivamente.

CONCLUSIONES

Del análisis estadístico realizado, se concluye lo siguiente:

1. SAG n°16

- El Tercer Período es el que reporta la mayor tasa de procesamiento fresco, con una diferencia positiva de 817.0579945 tpd respecto al Segundo Período, y en donde este último, tiene a su vez una diferencia positiva de 4573.0937055 tpd respecto al Primer Período. Cabe señalar que estos cálculos no consideran ni la disponibilidad ni la utilización de los equipos en los períodos señalados.
- El Tercer Período es el que reporta el mayor consumo de potencia, seguido del segundo y primer períodos respectivamente.

2. SAG n°17

- El Primer Período es el que reporta la mayor tasa de procesamiento fresco, con una diferencia positiva de 1304.8593785 tpd respecto al segundo período, y en donde este último, tiene una diferencia positiva de 4201.6379972 tpd respecto al tercer período. Cabe señalar que en estos cálculos no consideran ni la disponibilidad ni la utilización de los equipos en los períodos señalados.
- El Segundo Período es el que reporta el mayor consumo de potencia, seguido del tercer y primer períodos respectivamente.