Análisis de divergencia respecto al DANE

El objetivo de este documento es presentar una metodología para la comparación de los resultados del DANE y del DSI en torno al cálculo del valor agregado municipal.

Inicialmente, se importan las librerias necesarias:

```
library(readxl)
library(tidyverse)
library(factoextra)
library(scales)
library(ggplot2)
library(clustertend)
```

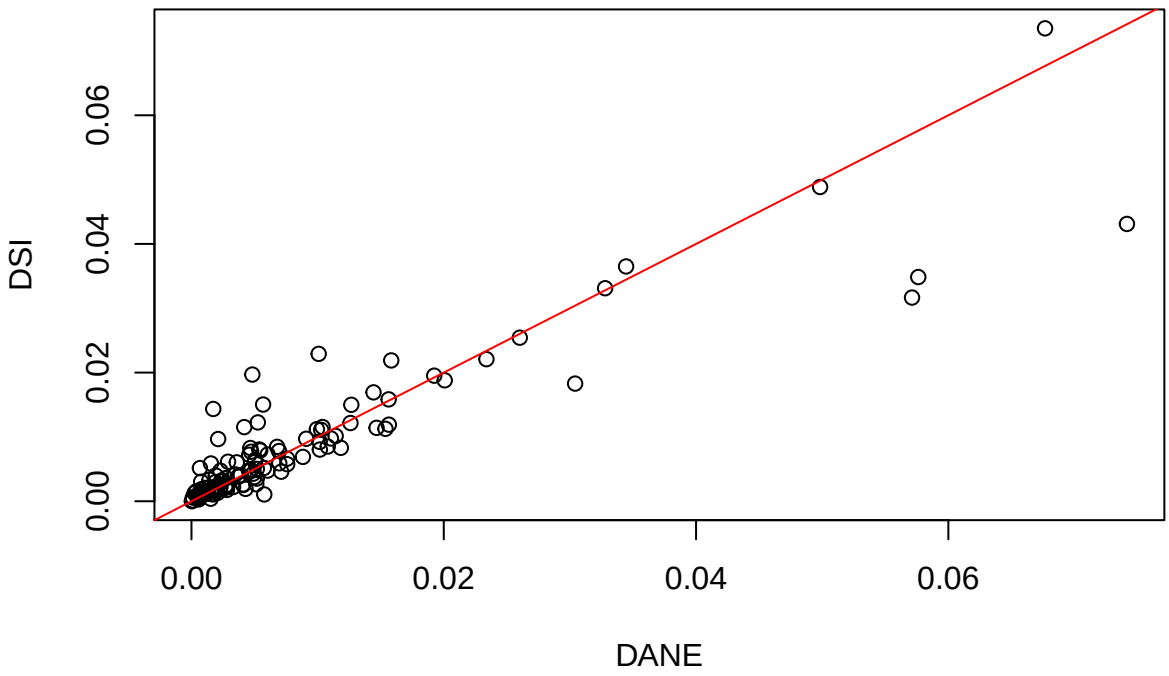
Posteriormente se importan los datos:

```
datos <- read_excel(file.choose())
```

Análisis a nivel sectorial

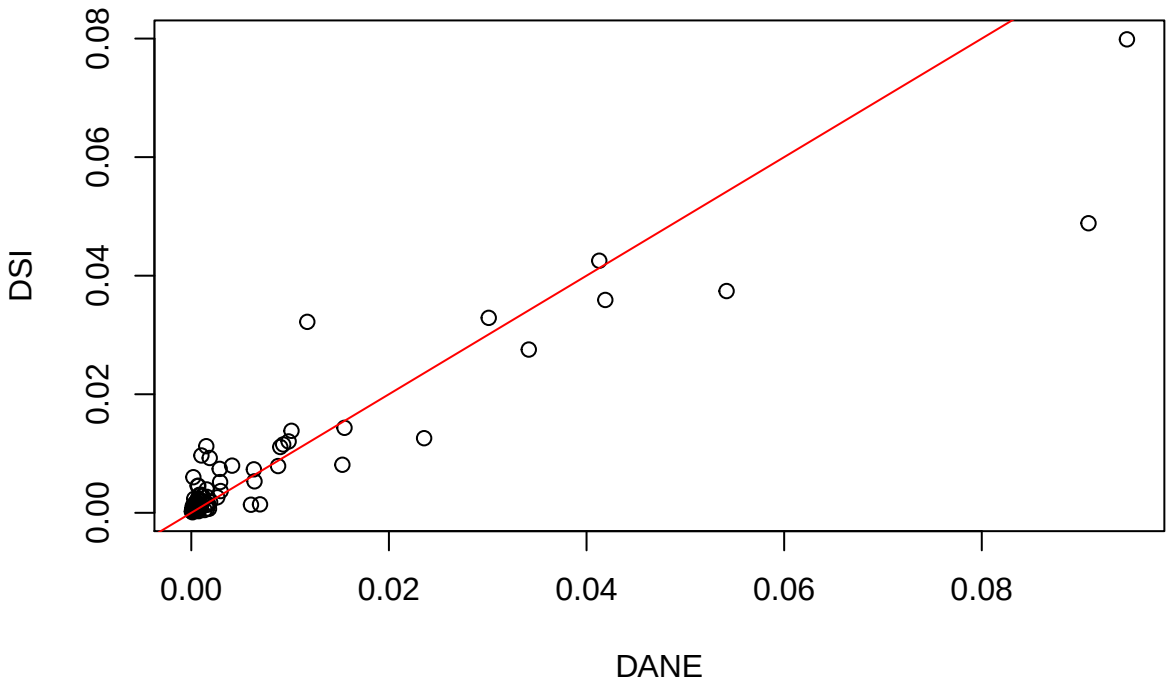
Las diferencias entre los resultados del DANE y el DSI para cada sector se pueden concebir como desviaciones de una recta de 45°. Para el primer sector, como puede observarse en el gráfico inferior, existe una concentración de los municipios con menor participación en torno a la línea de 45°, y las desviaciones más importantes a ese nivel se dan por sobreestimación de la participación respecto al valor reportado por el DANE. Cuando la participación de los municipios crece la desviación de la recta de 45° tiende a ser por una subestimación respecto a los valores reportados por el DANE.

```
a <- plot(datos$`S1 DANE`,datos$`S1 DSI`,xlab="DANE",ylab="DSI")+
abline(b=1,a=0,col="red")
```



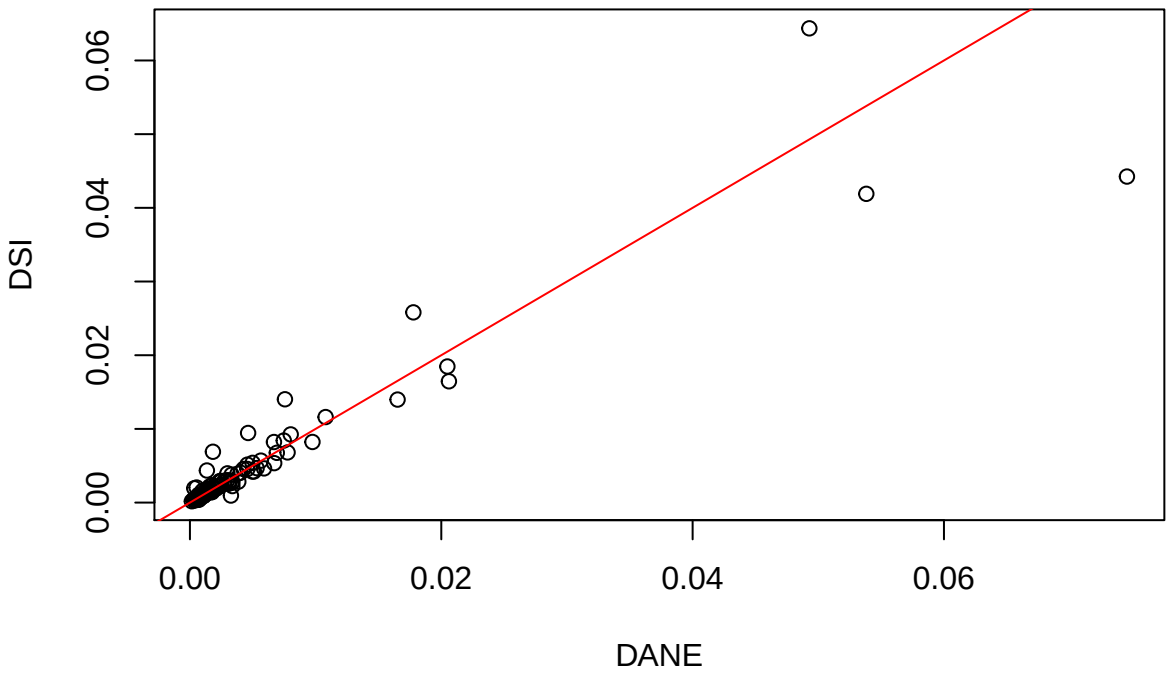
En el segundo sector, la concentración para participaciones inferiores es aún más acusada. Conforme aumentan las participaciones parece ser más prevalente la subestimación respecto al valor del DANE, pero a niveles más bajos prevalece la sobreestimación.

```
#En el gráfico se eliminan participaciones cercanas al 40% (Medellín)
#para comparación con los demás sectores.
b <- plot(datos$`S2 DANE`[datos$`S2 DANE`<0.4],
          datos$`S2 DSI`[datos$`S2 DSI`<0.3],
          xlab="DANE",ylab="DSI")+
abline(b=1,a=0,col="red")
```



El tercer sector tiene un comportamiento muy similar al segundo sector para las participaciones más bajas. A medida que aumentan las participaciones, las desviaciones tienden a tener un comportamiento similar por debajo y por encima, y tienden a abrirse.

```
c <- plot(datos$`S3 DANE`[datos$`S3 DANE`<0.4],
          datos$`S3 DSI`[datos$`S3 DANE`<0.4],
          xlab="DANE",ylab="DSI")+
abline(b=1,a=0,col="red")
```



De manera posterior a este acercamiento cualitativo, la pregunta de interés es cómo se comporta la distribución de las desviaciones para cada sector para obtener una noción cuantitativa de la magnitud de estas; en este caso la desviación se define al restar el valor de la participación obtenido por el DSI al valor del DANE. Para esto se hace un data frame con las columnas que corresponden a las diferencias, y se obtienen unas cuantas estadísticas descriptivas de las variables.

Como puede observarse en la información presentada en la parte inferior, el sector 1 tiene una desviación máxima 4 puntos porcentuales (en valor absoluto) y una desviación mínima de 0.0003%. El 75% de los datos se encuentra tiene una desviación menor al 0.23% respecto al valor del DANE.

El sector 2 es el sector con mayor dispersión respecto al promedio, que es ligeramente inferior al promedio del sector 1. Esto se debe a que, aunque tiene la desviación indidivual más alta de los tres sectores (4%), la mayor parte de las desviaciones son inferiores al 0.16%, de manera que la desviación de 4% es atípica, y las desviaciones se encuentran concentradas en un intervalo más estrecho para este sector (el 50% de ellas esta entre 0.02% y 0.16%) que para el sector 1.

```
diferencias <-datos[,c(2,9,10,11)]
names(diferencias) <- c("Mun","S1","S2","S3")
diferencias <- diferencias[order(diferencias$Mun),]
summary(diferencias[-1]*100)
```

```
##        S1                  S2                   S3
##  Min.   :-1.48699   Min.   :-2.047399   Min.   :-1.5070589
##  1st Qu.:-0.10768   1st Qu.:-0.102001   1st Qu.:-0.0232641
##  Median :-0.01118   Median :-0.032169   Median :-0.0005253
##  Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.0000000
##  3rd Qu.: 0.09385   3rd Qu.:-0.008337   3rd Qu.: 0.0150812
```

```
##  Max.   : 3.10534   Max.   : 4.195406   Max.   : 3.0316896
```

```
diferencias[,-1] %>% gather(key,value,S1:S3) %>% group_by(key) %>%
  summarise(min = min(abs(value)*100)
            ,Max = max(abs(value)*100)
            ,Media = mean(abs(value)*100)
            ,Desv = sd(abs(value)*100)
            ,Q25 = quantile(abs(value)*100, .25)
            ,Q75 = quantile(abs(value)*100, .75),
             .groups = 'drop')
```

```
## # A tibble: 3 x 7
##   key         min   Max Media  Desv     Q25    Q75
##   <chr>     <dbl> <dbl> <dbl> <dbl>   <dbl>  <dbl>
## 1 S1    0.000336   3.11 0.245 0.465 0.0322  0.237
## 2 S2    0.000961   4.20 0.215 0.490 0.0252  0.164
## 3 S3    0.0000712  3.03 0.111 0.338 0.00576 0.0605
```

¿La diferencia del 4% de este sector a qué municipio pertenece?

```
diferencias$Mun[abs(diferencias$S2*100)>4]
```

```
## [1] "Envigado"
```

Sin Envigado es interesante observar que la media desciende hasta el 0.18% y la desviación estándar a 0.33%, de manera que se puede afirmar que en el sector 2 las diferencias son más estrechas respecto al primer sector.

```
cat(paste(paste("La media es",
                mean(abs(diferencias$S2*100)[abs(diferencias$S2*100)<4])),
          paste("La desviación es",
                sd(abs(diferencias$S2*100)[abs(diferencias$S2*100)<4])),
          paste("El nuevo valor máximo es",
                max(abs(diferencias$S2*100)[abs(diferencias$S2*100)<4])),
          sep="\n")
    )
```

```
## La media es 0.183267818289074
## La desviación es 0.335583219299385
## El nuevo valor máximo es 2.04739936255786
```

En el caso del sector 3, el valor máximo está apenas por debajo del valor máximo del sector 1, y el valor mínimo es el menor de todos. El 75% de los datos está por debajo de una diferencia absoluta de 0.06%. En este sector, Envigado vuelve a surgir como un municipio donde la diferencia es más elevada que el resto; sin este valor, las estadísticas descriptivas son las menores de todos los sectores. Se puede afirmar que en este sector, el acercamiento es el más preciso excepto para Envigado.

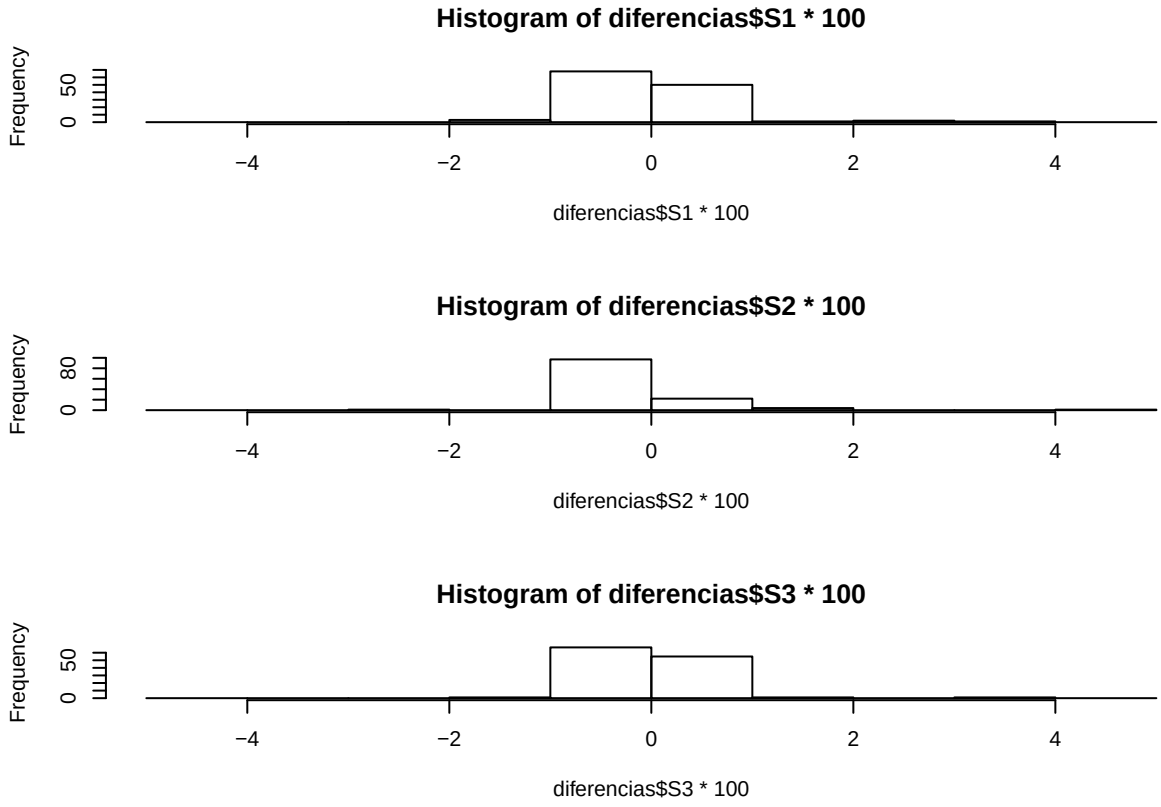
```
cat(paste(paste("La media es",mean(abs(diferencias$S3*100)[abs(diferencias$S3*100)<3])),
          paste("La desviación es",
sd(abs(diferencias$S3*100)[abs(diferencias$S3*100)<3])),
paste("El nuevo valor máximo es", max(abs(diferencias$S3*100)[abs(diferencias$S3*100)<3])),
sep="\n"))
```

```
## La media es 0.0871571619186016
## La desviación es 0.213154571105213
## El nuevo valor máximo es 1.50705887656838
```

Dejando de lado la caracterización por diferencias absolutas, ¿existe un sesgo a la subestimación o sobreestimación?

Visualizando la distribución de los datos mediante un histograma, se observa que la mayor parte de los datos tiene una diferencia negativa, es decir, existe un sesgo a la sobreestimación. El tercer sector es el que está más homogéneamente distribuido, y el segundo sector es el que presenta este rasgo de manera más acusada.

```
par(mfrow=c(3,1))
hist(diferencias$S1*100,breaks=c(-5,-4,-3,-2,-1,0,1,2,3,4,5))
hist(diferencias$S2*100,breaks=c(-5,-4,-3,-2,-1,0,1,2,3,4,5))
hist(diferencias$S3*100,breaks=c(-5,-4,-3,-2,-1,0,1,2,3,4,5))
```



Para todos los sectores, las diferencias son más densas en el intervalo del 1% por debajo y por encima. Los municipios que están por fuera de estos valores para el primer sector son:

```
cat(
  paste(
    paste("Los municipios con diferencias superiores al 1% son:",
          paste(diferencias$Mun[diferencias$S1*100>1],
                collapse=", ")),
    paste("Los municipios con diferencias inferiores al 1% son:",
          paste(diferencias$Mun[diferencias$S1*100<c(-1)],
```

```
                collapse=", ")),
    sep="\n"
  )

)
```

```
## Los municipios con diferencias superiores al 1% son: Apartadó, Carepa, Chigorodó, Turbo
## Los municipios con diferencias inferiores al 1% son: Barbosa, El Carmen de Viboral, El Santuario
```

Para el segundo sector:

```
cat(
  paste(
    paste("Los municipios con diferencias superiores al 1% son:",
          paste(diferencias$Mun[diferencias$S2*100>1],
                collapse=", ")),
    paste("Los municipios con diferencias inferiores al 1% son:",
          paste(diferencias$Mun[diferencias$S2*100<c(-1)],
                collapse=", ")),
    sep="\n"
  )

)
```

```
## Los municipios con diferencias superiores al 1% son: Envigado, Guarne, Itagüí, Medellín, Rionegro
## Los municipios con diferencias inferiores al 1% son: Barbosa
```

Para el tercer sector:

```
cat(
  paste(
    paste("Los municipios con diferencias superiores al 1% son:",
          paste(diferencias$Mun[diferencias$S3*100>1],
                collapse=", ")),
    paste("Los municipios con diferencias inferiores al 1% son:",
          paste(diferencias$Mun[diferencias$S3*100<c(-1)],
                collapse=", ")),
    sep="\n"
  )

)
```

```
## Los municipios con diferencias superiores al 1% son: Envigado, Itagüí
## Los municipios con diferencias inferiores al 1% son: Bello
```

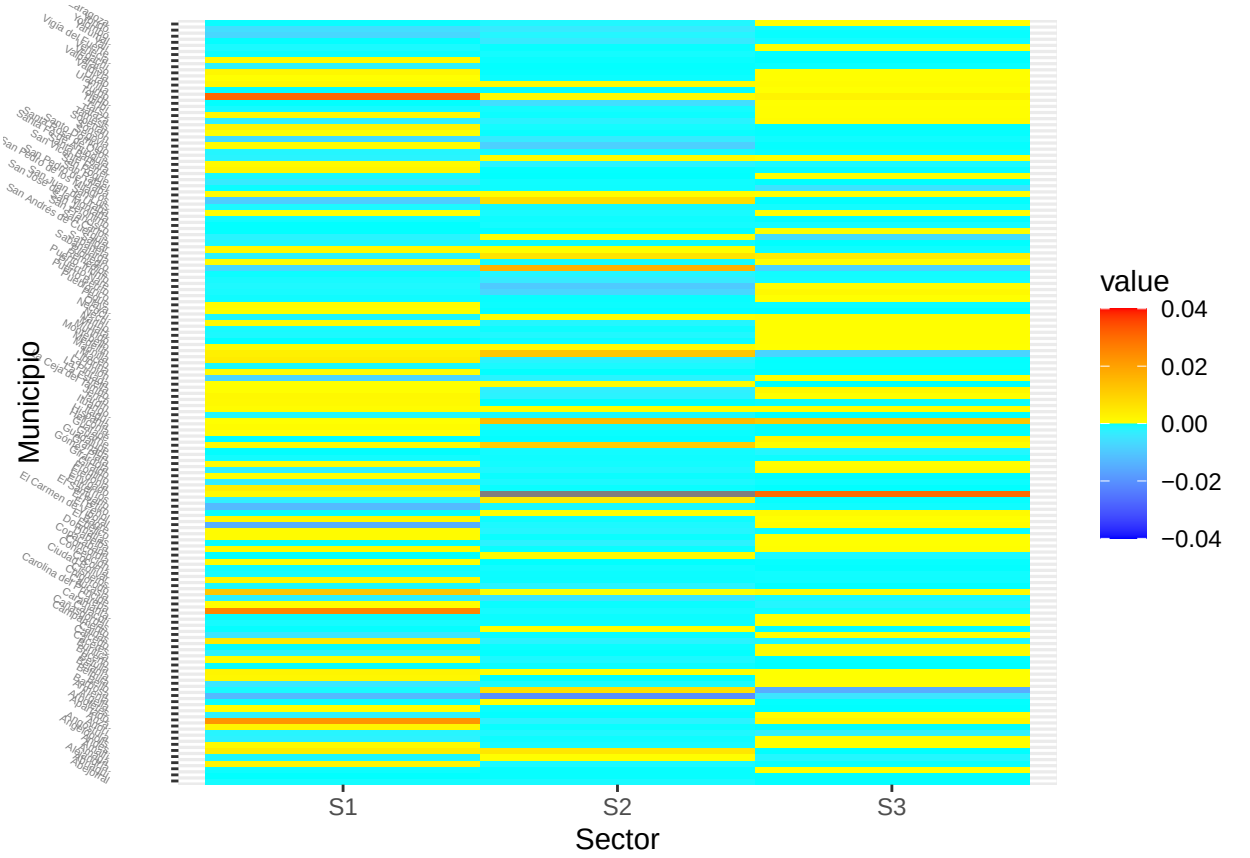
Las principales conclusiones son entonces:

1. Existe un sesgo a la sobreestimación por parte del DSI.
2. La mayoría de las diferencias son inferiores al 1% en valor absoluto, siendo el primer y el segundo sector los que tienen más municipios por encima de ese valor

3. El tercer sector es el sector donde la estimación es más precisa; la mayor parte de los datos se encuentra en un intervalo más estrecho en torno al cero que los demás sectores; su media es la menor (sin Envigado) y su dispersión es de las más bajas. En el primer sector ocurre lo contrario.

Este análisis se puede resumir en un mapa de calor. El siguiente código permite realizar un mapa de calor que es exportado como una imagen externa debido a su tamaño. Se recomienda ejecutarlo para visualizar la imagen.

```
diferencias %>%
        gather(key,value,S1:S3) %>%
        ggplot(aes(x = factor(key), y = factor(Mun), fill = value)) +
        geom_raster() +
        labs(x = "Sector", y = "Municipio")+
        theme(axis.text.y = element_text(size = 4, angle =330, colour ="grey50"))+
        scale_fill_gradientn(colours=c("blue","cyan","white", "yellow","red"),
                             values=rescale(c(-0.04,0-.Machine$double.eps,0,0+.Machine$double.eps,0.04)),
                             limits = c(-0.04, 0.04))
```




```
ggsave("heatmap.png",dpi="retina",height=8,width=3)
```

¿Es posible agrupar estos municipios en clusters? Para comprobar si existe suficiente separación en los municipios a lo largo de las tres dimensiones (las diferencias en los tres sectores) se hace la prueba de Hopkins de tendencias de cluster.

Según Kassambara (2017), en esta prueba la hipótesis nula consiste en que no existen clusters o grupos claramente separables entre los datos, y la hipótesis nula es su contrario. Para evaluar estas hipótesis se

calcula el estadístico H, el cual, si tiene un valor cercano a 0.5 implica la no existencia de clusters y visceversa. Procedemos a calcular dicho estadístico:

```
hopkins(diferencias[-1],n=nrow(diferencias)-1)
```

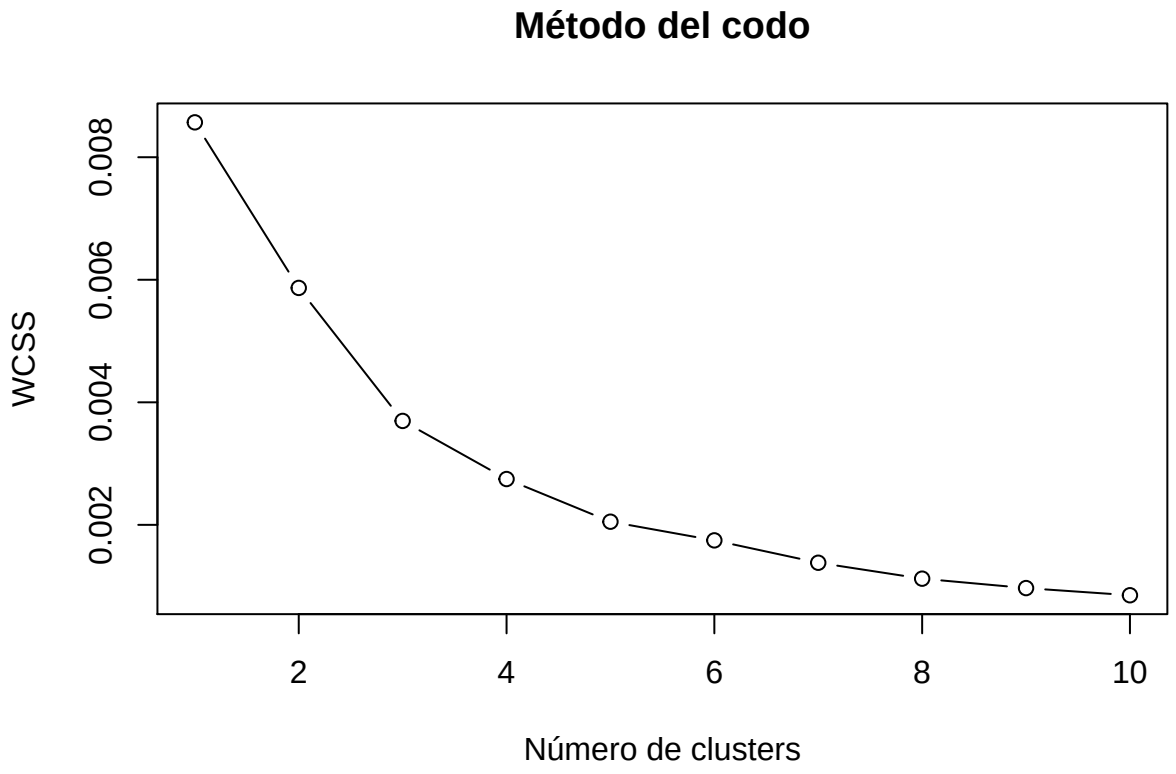
```
## $H
## [1] 0.1075448
```

Ya que el valor es inferior a 0.5, se concluye que existen clusters en los datos.

Kassambara (2017) también presenta un algoritmo de clustering que permite obtener soluciones únicas, y que solucionan el problema de sensibilidad a los datos iniciales que presenta el algoritmo K-Medias. Este algoritmo combina el algoritmo de cluster jerárquico y el algoritmo K-Medias, y al igual que el último, requiere la especificación del número de clusters presentes en los datos. Esto lo obtenemos mediane el método del codo.

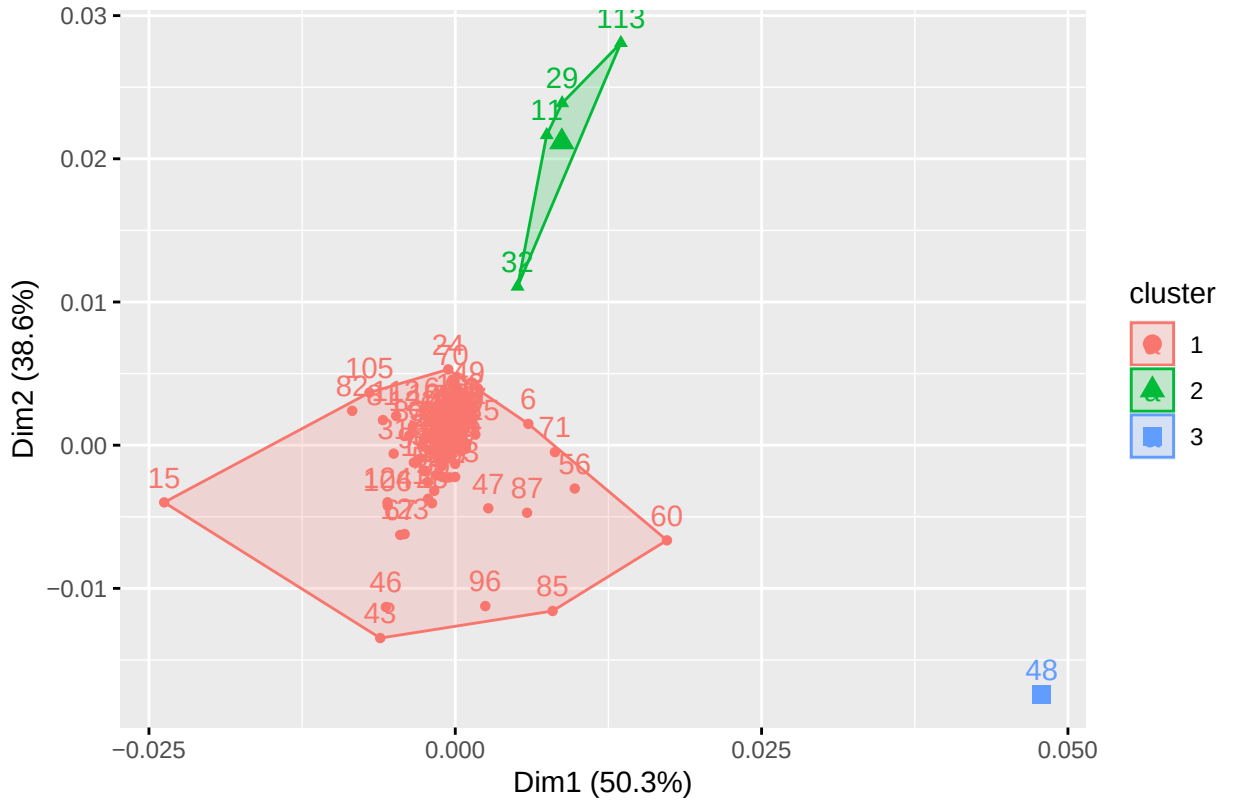
En el gráfico inferior puede observarse que el codo del gráfico o el punto para el cual la pendiente muestra una reducción significativa (en valor absoluto, es decir, se acerca al cero) es k=3.

```
wcss = vector()
for (i in 1:10) wcss[i] <- sum(hkmeans(datos[,c(9,10,11)], i)$withinss)
plot(1:10,
     wcss,
     type = 'b',
     main = paste('Método del codo'),
     xlab = 'Número de clusters',
     ylab = 'WCSS')
```



Se procede a calcular el algoritmo para tres clusters. En el gráfico inferior podemos observar que los grupos están claramente separados entre sí. Los ejes corresponden a los dos primeros componentes principales que dan cuenta de cerca del 89% de variación en los datos; esta técnica se usa con el fin de reducir la dimensión de los datos de tres a dos para capturar los grupos en un gráfico bidimensional y la aplica de manera automática la función fviz_cluster().

```
y <- hkmeans(datos[,c(9,10,11)],k=3)
fviz_cluster(y,data=datos[,c(9,10,11)],stand = FALSE,main="")
```



En la tabla inferior, se puede observar que la mayoría de los municipios pertenece al primer cluster. Este se caracteriza por tener en promedio sobreestimaciones en todos los sectores, de 0.07%, 0.03% y 0.3% respectivamente. En este cluster parecen estar todos los municipios que se ubican en la base de las rectas de 45° de los tres primeros gráficos de este documento.

El segundo cluster se caracteriza por tener en promedio, subestimaciones respecto al valor del DANE en el sector primario y secundario, siendo mayor la del sector primario, y una sobreestimación en el sector secundario del 0.04%.

Finalmente, el tercer cluster contiene municipios que, en promedio, subestiman el valor del DANE; en particular, lo subestiman en valores extremos en los sectores 2 y 3.

```
y$centers*100
```

```
##   S1 DANE-DSI S2 DANE-DSI S3 DANE-DSI
## 1  -0.0771143 -0.03343417 -0.02985202
## 2   2.2850401 -0.04582645  0.13763806
## 3   0.1135557  4.19540572  3.03168961
```

¿Qué municipios conforman el segundo y tercer cluster? Son algunos de los municipios que se identificaron anteriormente que se alejan del resto en los histogramas. El tercer cluster lo conforma solamente Envigado, el cual mostraba gran distancia del resto de los datos para los sectores 2 y 3.

```
cat(paste(
  paste("Segundo cluster:",paste(diferencias$Mun[y$cluster==2],collapse=", ")),
  paste("Tercer cluster:",paste(diferencias$Mun[y$cluster==3],collapse=", ")),
  sep="\n"
  )
)
```

```
## Segundo cluster: Apartadó, Carepa, Chigorodó, Turbo
## Tercer cluster: Envigado
```

Si se trata el valor del DANE como el valor de referencia y el valor del DSI como un estimado de dicho valor, es posible cuantificar la desviación como si se tratara del análisis del ajuste de un modelo de predicción y usar una función de pérdida. En este caso se escoge la función RMSE (raíz cuadrada del error cuadrático medio); los valores arrojados por la función tienen sentido solo en términos relativos.

El resultado obtenido confirma la conclusión obtenida previamente: en el tercer sector se verifica una menor desviación respecto al valor del DANE; en el segundo sector, el resultado obtenido se explica probablemente por el efecto de Envigado.

```
RMSE = function(m, o){
  sqrt(mean((m - o)^2))
}

cat(paste(
  paste("El RMSE del primer sector es:", RMSE(datos$`S1 DSI`,datos$`S1 DANE`)),
  paste("El RMSE del segundo sector es:", RMSE(datos$`S2 DSI`,datos$`S2 DANE`)),
  paste("El RMSE del tercer sector es:",RMSE(datos$`S3 DSI`,datos$`S3 DANE`)),
  sep="\n"
  )
   )
```

```
## El RMSE del primer sector es: 0.00524237625941628
## El RMSE del segundo sector es: 0.00533801526109283
## El RMSE del tercer sector es: 0.00354644397069697
```

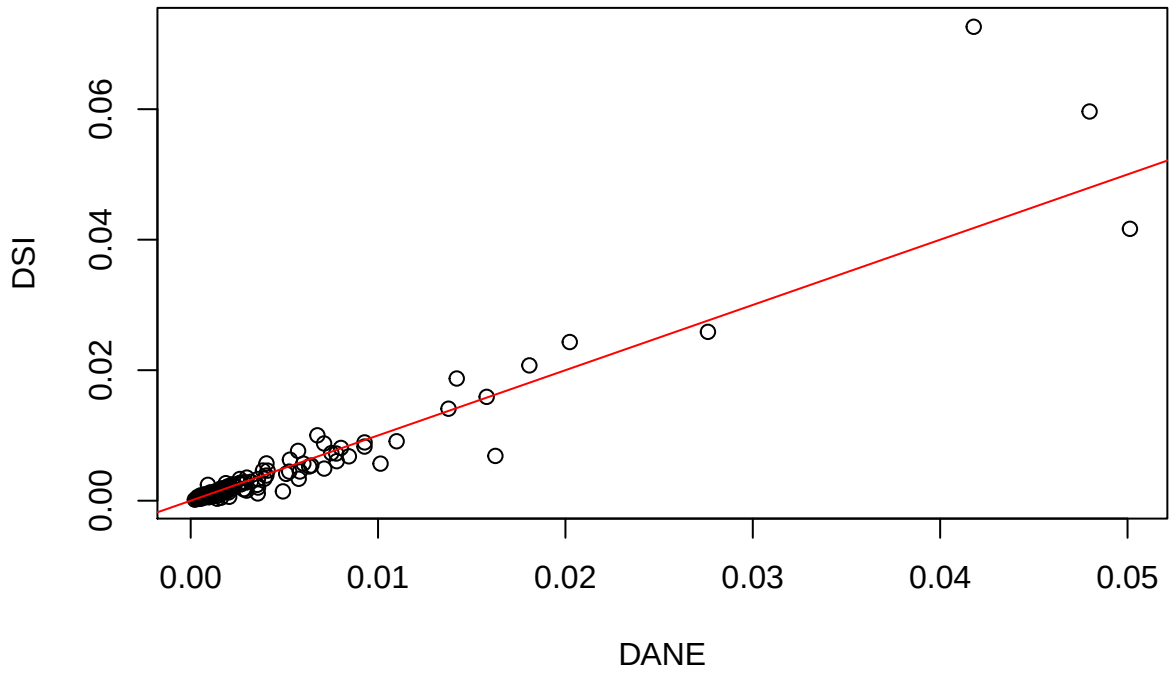
Análisis a nivel agregado

Inicialmente, se agregan los valores reportados por sector y se hallan las diferencias:

```
datos["DANE_VA"] <- apply(as.matrix(datos[,12:14]),1,sum)/sum(apply(as.matrix(datos[,12:14]),1,sum))
datos["DSI_VA"] <- apply(as.matrix(datos[,15:17]),1,sum)/sum(apply(as.matrix(datos[,15:17]),1,sum))
datos["DIF_VA"] <- datos["DANE_VA"]-datos["DSI_VA"]
```

Al observar las participaciones para el valor agregado por municipio, se observa que la mayoría es inferior a 1% en valor absoluto y que para participaciones más grandes, la desviación es más elevada. No es claro en este punto si hay sesgo a subestimar o sobrestimar; a valores más pequeños parece haber mayor densidad de valores que subestiman, y a mayor participación, parece que se tiende a sobreestimar.

```r
plot(datos$DSI_VA[datos$DSI_VA<0.4],datos$DANE_VA[datos$DANE_VA<0.4],xlab="DANE",ylab="DSI")+
abline(b=1,a=0,col="red")
```




```
## integer(0)
```

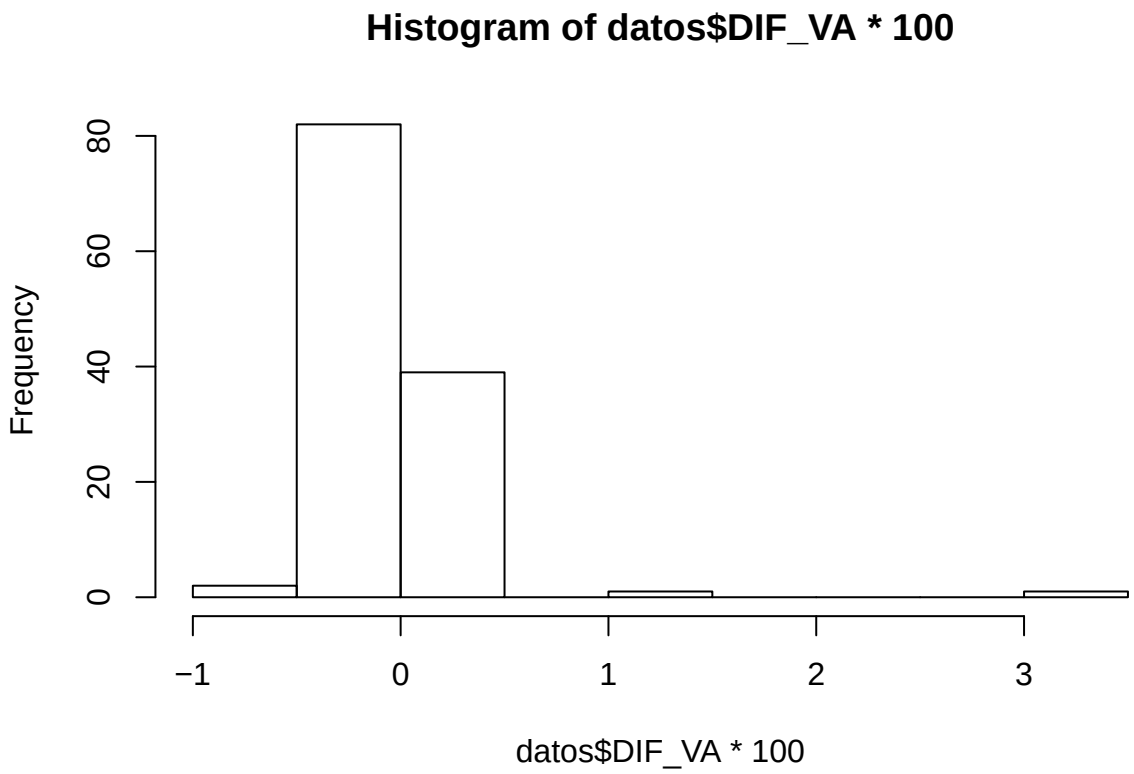
La desviación máxima corresponde al 3% respecto a la participación estimada por el DANE, y la mínima del 0.0001%. El 75% de los datos tiene una desviación inferior al 0.1%.

```r
summary(abs(datos$DIF_VA*100))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000977 0.0106019 0.0317911 0.1146251 0.1072821 3.0834936
```

Visualizando el histograma queda más claro que en la participación del valor total se tiene un sesgo a la sobreestimación. La desviación para la mayoría de municipios está por debajo del 1% en valor absoluto.

```r
hist(datos$DIF_VA*100)
```



Los dos municipios que tienen una desviación en la participación superior al 1% son:

```
datos$MUNICIPIO[datos$DIF_VA*100>1]
```

```
## [1] "Envigado" "Itagüí"
```

El RMSE para la desviación de la participación en el valor total es:

```
RMSE(datos$DSI_VA,datos$DANE_VA)
```

```
## [1] 0.003354376
```