Bitácora 2

Ordenamiento de la literatura

Table 1: Ordenamiento de la literautra

Tipo de grupo	**Nombre del grupo**	**Nombre del tema**	**Título**	**Año**	**Autor (es)**
Metodológico	Modelo de cópulas	Modelación de distribuciones de pérdidas	Modelling Dependencies in Airport Passenger Claim Data Using Copulas	2022	Roberto Carcache Flores
Metodológico	Modelos de pérdidas agregadas	Modelación paramétrica de distribuciones de pérdidas	Aggregate loss model with Poisson - Tweedie loss frequency	2020	Si Chen
Metodológico	Estimación de densidades por kernels	Modelación no paramétrica de distribuciones de pérdidas	Estimation of Parametric and Non-parametric Models for Univariate Claim	2011	David Pitt, Montserrat Guillen y Catalina Bolancé
Metodológico	Modelos de frecuencia y severidad	Modelación de las distribuciones de frecuencia y severidad	Severity Distributions - an approach using R	2018	Cyprian Ondieki, Shalyne Gathoni y Joan Wairimu

Enlaces de la literatura

En @Flores se establece el procedimiento base para conseguir las distribución agregada al igual que algunos hallazgos y metodologías que son de alta utilidad. Primero la agregación de los datos se hace mensualmente con suma para la severidad y por frecuencia para los reclamos. El autor nota que hay un tendencia negativa de la frecuencia y severidad con respecto al tiempo por lo que procede a eliminarla. Luego, determina la mejor distribución para cada variable utilizando estimación de máxima verosimilitud (MLE). Se encuentra que la binomial negativa se ajusta mejor a las frecuencias. Por otro lado, la Log-Laplace se ajusta mejor a los reclamos por daños a la propiedad y la lognormal se ajusta mejor a los reclamos por pérdidas de los bienes, por lo que se utilizan estas dos para modelar la severidad. Durante este proceso el autor nota que la eliminación de la tendencia facilita el proceso de ajustar una distribución a la frecuencia y la severidad. Finalmente, las cópulas multivariadas se comparan utilizando log verosimilitud y se obtiene que las cópulas elípticas (Gaussiana y t-Student) se ajustan mejor que las arquimedianas (Clayton y Gumbel).

En un estudio similar, @pitt2011estimation utilizan datos de costos de reclamos hechos a una aseguradora española por accidentes ocurridos en el año 2000 y recopilados en 2002, que incluye tanto los ligados a costos por daños a la propiedad como por costos médicos. El tamaño de muestra es de 518 reclamos. Al igual que el estudio anterior, para estimar la densidad para cada uno de los costos (daños a la propiedad y médicos) se utilizan métodos paramétricos como las aproximaciones normales y log-normales. En contraste al estudio pasado también recurren a estimadores no paramétricos como la aproximación por kernels modificada, donde la modificación consiste en que primero se aplica una transformación a los datos originales para corregir la asimetría, se hace una aproximación con un kernel gaussiano a los datos modificados, y luego se calcula la aproximación de los datos originales a partir de la calculada para los modificados. La transformación aplicada a los datos se enmarca en la *shifted power transformation family*.

Adicionalmente, los mismos autores exponen métodos para evaluar la bondad de ajuste de las distribuciones encontradas. Para evaluar todas las estimaciones propuestas se utiliza la log-verosimilitud, tanto la versión clásica como modificaciones ponderadas, mientras que para evaluar solamente los métodos no paramétricos se usan distintas versiones de una aproximación a errores cuadráticos integrados ponderados. Se concluye que la log-verosimilitud no es una buena medida de bondad de ajuste para comparar los ajustes no paramétricos, debido a su relación inversa con la magnitud del ancho de banda empleado. En general, de las propuestas paramétricas, la log-normal tuvo un mejor desempeño, el cual es un hallazgo que concuerda con el de @Flores, mientras que la estimación por kernel modificada tuvo un desempeño adecuado y se recomienda para modelar distribuciones con colas pesadas.

Tablas

Cabe mencionar que después de eliminar observaciones con datos desconocidos (NA), se cuenta con un total de 94 848 observaciones.

```
medidas <- function(x){
  r <- summary(x) %>% as.vector()
  temp <- data.frame(c(r, sd(x), IQR(x),skewness(x), kurtosis(x)))
  return(temp)
}

nombres <- c("Mínimo",
             "Primer cuartil",
             "Mediana",
             "Media",
             "Tercer cuartil",
             "Máximo",
             "Desviación estándar",
             "Rango intercuartil",
             "Asimetría",
             "Curtosis")

tabla <- medidas(close_amount)
rownames(tabla) <- nombres

tabla %>%
  kbl(caption = "Medidas de resumen", col.names = F) %>%
  kable_styling() %>%
  kable_classic_2(full_width = T) %>%
   row_spec(0, bold = T)
```

Se cargan los datos al estilo David, no entiendo cual es el problema con hacerlo asi

```
df1 <- datos %>% group_by( month(date_received) ) %>%
  summarise(es = mean(close_amount, na.rm = TRUE))
colnames(df1) <- c("Mes", "Severidad promedio")

kable(df1) %>% kable_styling(full_width = FALSE)
```

Severidad promedio por mes de los reclamos

Table 2: Medidas de resumen

	FALSE
Mínimo	0.00000
Primer cuartil	0.00000
Mediana	0.00000
Media	89.04626
Tercer cuartil	61.97250
Máximo	250000.00000
Desviación estándar	882.05884
Rango intercuartil	61.97250
Asimetría	241.96787
Curtosis	68021.83944

Mes	Severidad promedio
1	62.78637
2	66.81869
3	64.63042
4	65.62982
5	67.52037
6	67.92822
7	63.20689
8	59.25496
9	59.79396
10	59.67962
11	58.63796
12	65.79952
NA	56.84959

Tipo de reclamo	Ocurrencias
	7913
-	282
Bus Terminal	1
Complaint	48
Compliment	3
Employee Loss (MPCECA)	485
Motor Vehicle	369
Passenger Property Loss	117868
Passenger Theft	465
Personal Injury	1465
Property Damage	75364
Wrongful Death	4

```
df2 <- datos %>% group_by( claim_type ) %>% summarise(n = n())
colnames(df2) <- c("Tipo de reclamo", "Ocurrencias")

kable(df2) %>% kable_styling(full_width = FALSE)
```

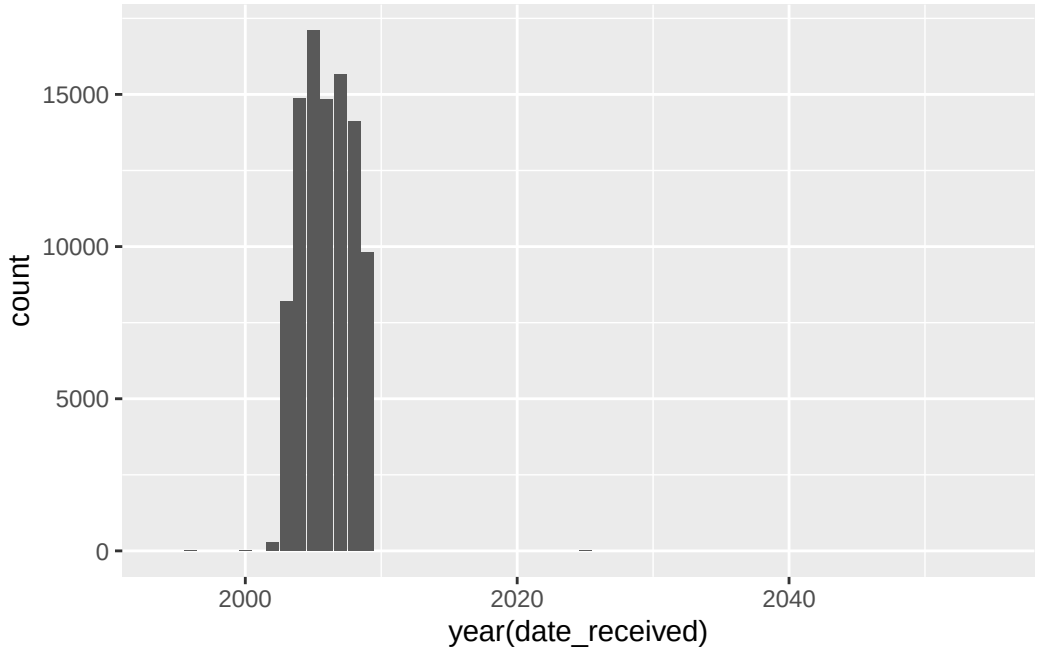
Conteo reclamos por tipo

Gráficos

```
a <- min(year(date_received))
b <- max(year(date_received))

ggplot(base, aes(year(date_received))) +
  stat_count(geom='bar', aes(y=..count..))+
  scale_x_continuous(limits = c(a,b))
```

```
Warning: Removed 2 rows containing missing values (geom_bar).
```




```
df <- datos %>% group_by(year(date_received), month(date_received) ) %>%
  summarise(s = sum(close_amount, na.rm = TRUE), n=n())
```

```
`summarise()` has grouped output by 'year(date_received)'. You can override
using the `.groups` argument.
```

```
df <- cbind( t=1:179, df)
df <- df[1:172,]
```

```
ggplot(df, aes( x=t, y = n)) + geom_point(color='red', size=2) +
  xlab("Tiempo") + ylab("Reclamos")+
  theme_minimal()
```

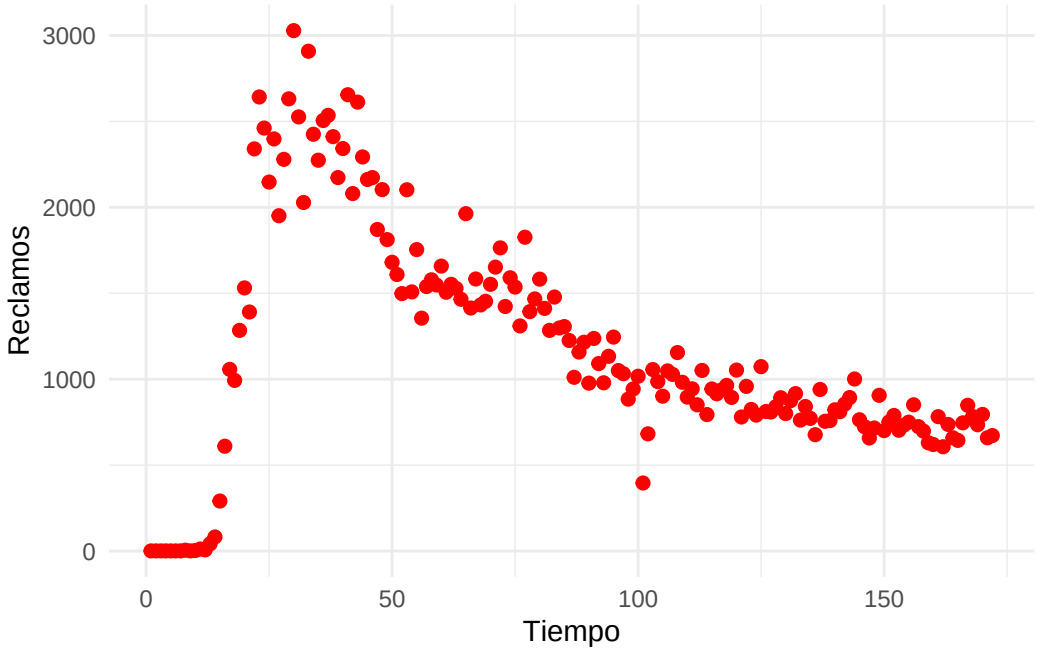


Figure 1: Número de reclamos mensuales del 2002 al 2015

```
df3 <- datos %>% group_by(year(date_received), month(date_received) ) %>%
  summarise(s = mean(close_amount, na.rm = TRUE)) %>% na.omit()
```

```
`summarise()` has grouped output by 'year(date_received)'. You can override
using the `.groups` argument.
```

```
df3 <- cbind(t=1:105, df3)
```

```
ggplot(df3, aes( x=t, y = s)) + geom_point(color='red', size=2) +
  xlab("Tiempo") + ylab("Severidad promedio")+
  theme_minimal()
```

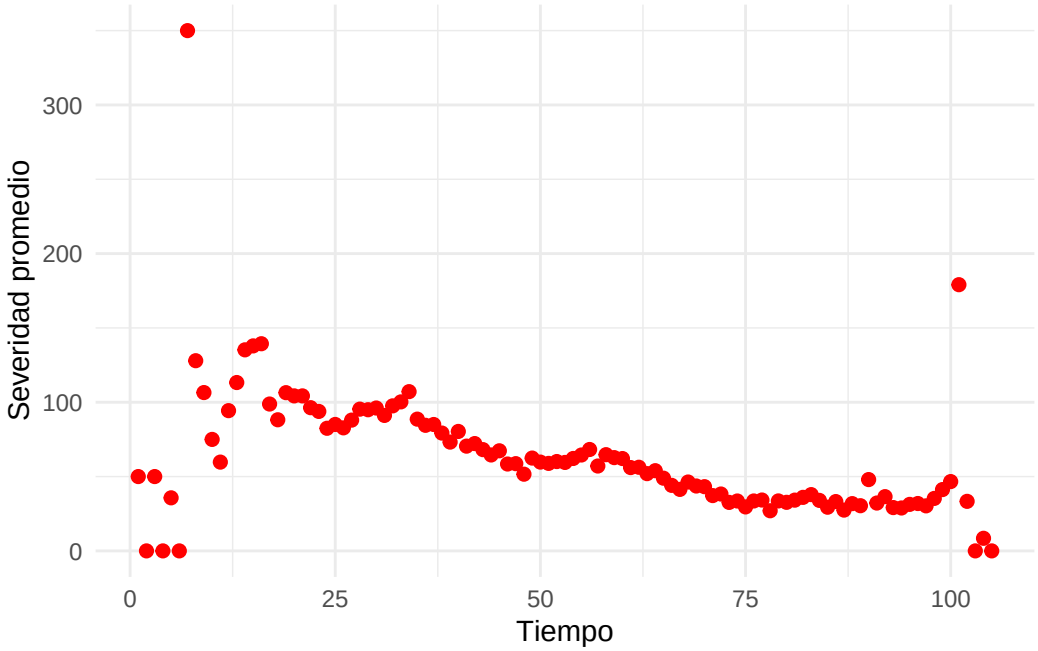
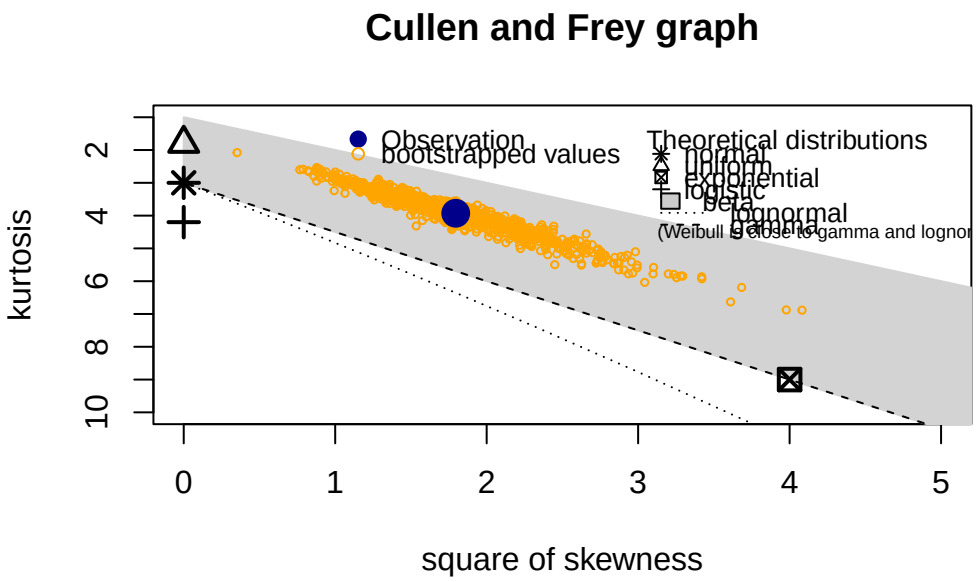


Figure 2: Severidad promedio mensual entre 2002 y 2015

```
descdist( df$s, discrete=FALSE, boot=1000)
```




```
summary statistics
------
min:  0   max:  264842.8
median:  17222.44
mean:  49120.29
estimated sd:  65562.32
estimated skewness:  1.34006
estimated kurtosis:  3.928245
```

Fichas Bibliográficas:

- Nombre de su hallazgo/resultado: Tendencia negativa de los reclamos
- Resumen en una oración: El número de reclamos mensuales parece hacer incrementado rápidamente poco tiempo de abrirse el TSA hasta alcanzar y máximo y desde entonces se ha mostrado un comportamiento a la baja de la cantidad de reclamos hechos.
- Principal característica: Tendencia negativa
- Problemas o posibles desafíos: En @Flores, se comenta que la existencia de una tendencia en los reclamos puede causar problemas al momento de buscar las distribuciones que se ajusten a los datos.

- Resumen en un párrafo: El número de reclamos mensuales parece hacer incrementado rápidamente poco tiempo de abrirse el TSA. Esto se podría explicar por la poca experiencia en materia de chequeos y procedimientoslo que posiblemente causó problemas mala práctica con los pasajeros. Sin embargo, luego de alcanzarse un máximo, este comportamiento cambia a un decrecimiento desde entonces. Esto se debe a que probablemente el TSA se ha vuelto mejor con el manejo de los chequeos. Esta tendencia puede ser un problema porque en la literatura se expresó que puede complicar el proceso de ajustar una distribución a los reclamos, notando que al eliminar esta tendencia se facilitaba esta búsqueda.