Análisis espacio-temporal de patrones puntuales sobre la Guerra de Afganistán

Análisis estadístico-predictivo de datos para prevención de conflictos

Sandra Martín-Forero Cogolludo

www.linkedin.com/in/sandra-mfc

Sandra.mfc18@gmail.com

Índice

1. *Introducción*

Este proyecto se basa en un dataset que contiene alrededor de 77000 eventos los cuales representan desde episodios centrados en detención y registro de propiedades hasta tiroteos, desde el año 2004 hasta el año 2009. Cada evento representa un punto en el mapa, y cabe destacar que hay que tener en cuenta que estos episodios puntuales se reducen a 75239 eventos una vez que ya se han excluido los puntos que se sitúan fuera de la frontera de Afganistán.

El principal objetivo se centrará en buscar de manera global para dichos puntos algún patrón identificable en función del tiempo, teniendo en cuenta sus intensidades, gracias a tres metodologías distintas: Scott, Cronie y Van Lieshout y Validación Cruzada, además de intentar responder a una compleja cuestión: ¿es posible predecir una guerra?

A su vez, la función K nos indicará si se forman clústers, o de lo contrario, si habrá distribuciones dispersas. **La función K** es el número de eventos que ocurren en un radio r alrededor de cualquier otro evento; esto significa que la función K **representa el número medio de eventos dentro de un círculo de radio r alrededor de un evento típico del patrón.**

Figura 1.1. Mapa de la representación de eventos en la escalada del conflicto de Afganistán

Como puede apreciarse, sin colores resulta algo caótico y no se pueden extraer demasiadas conclusiones; aquí radica la importancia de este estudio. Una vez desarrollado, se podrá comprobar con claridad cuáles son los puntos clave del conflicto, al igual que los meses de la escalada.

Cabe destacar que para este trabajo se hace uso de la Estadística Espacial y Espacio-Temporal porque es la rama de la estadística que mejor interpreta los patrones puntuales de eventos que ocurren en el tiempo. Los procesos puntuales (que es en lo que se basa este trabajo), son modelos estocásticos complejos que se basan en la descripción de la localización de eventos de interés.

2. *Objetivos del trabajo:*

- Obtener un patrón en los datos que establezca cuándo ocurre la escalada del conflicto.
- Comparar los resultados obtenidos para ver la diferencia en las tres metodologías distintas para obtener los bandwidth: ppl, CvL y Scott.
- Comparar los resultados obtenidos usando las medias aritméticas y las medias geométricas a la hora de obtener un único bandwidth.
- Establecer las zonas de Afganistán donde ocurren mayor número de eventos como los descritos.
- Obtener las funciones K para representar la interacción entre los puntos y poder estudiar los cambios en función del tiempo.

3. *Estructura del trabajo:*

Siguiendo la metodología CRISP-DM, **la primera fase** que describe como la **'comprensión del negocio',** es donde se exponen los objetivos que se persiguen desarrollando este estudio, se evalúa la situación y se adapta el proyecto a las necesidades de los datos que se tienen.

En la segunda fase (comprensión de los datos) se plasma la idea general de las conclusiones que se buscan obtener a partir del data set proporcionado. Los datos se recogen en un data set formado por 72 elementos; y cada elemento del mismo contiene unas coordenadas que proporcionan el evento puntual que corresponde.

La tercera fase englobaría la **obtención de los bandwidth,** el **cálculo** de las medias aritmética y geométrica, las funciones de densidad y la función K, mientras que **la cuarta fase (modelado)** se relaciona con el ploteo de los datos calculados.

La quinta fase se refiere a la **evaluación de los resultados:** aquí se obtendrán las conclusiones extraídas del estudio realizado. Y para finalizar, en **la sexta fase** (el final de este proyecto) **se decidirá qué método se adapta mejor** a los datos y se implementará en función de los mismos.

4. *Entorno y análisis:*

En la figura 2.1. se muestra cómo evoluciona la escalada en función de los meses, desde enero de 2004 hasta diciembre de 2009. Los colores indican dónde se concentra el mayor número de eventos, de lo que puede deducirse a simple vista que según este método, en agosto de 2009 es donde se produce el punto álgido de la escalada.

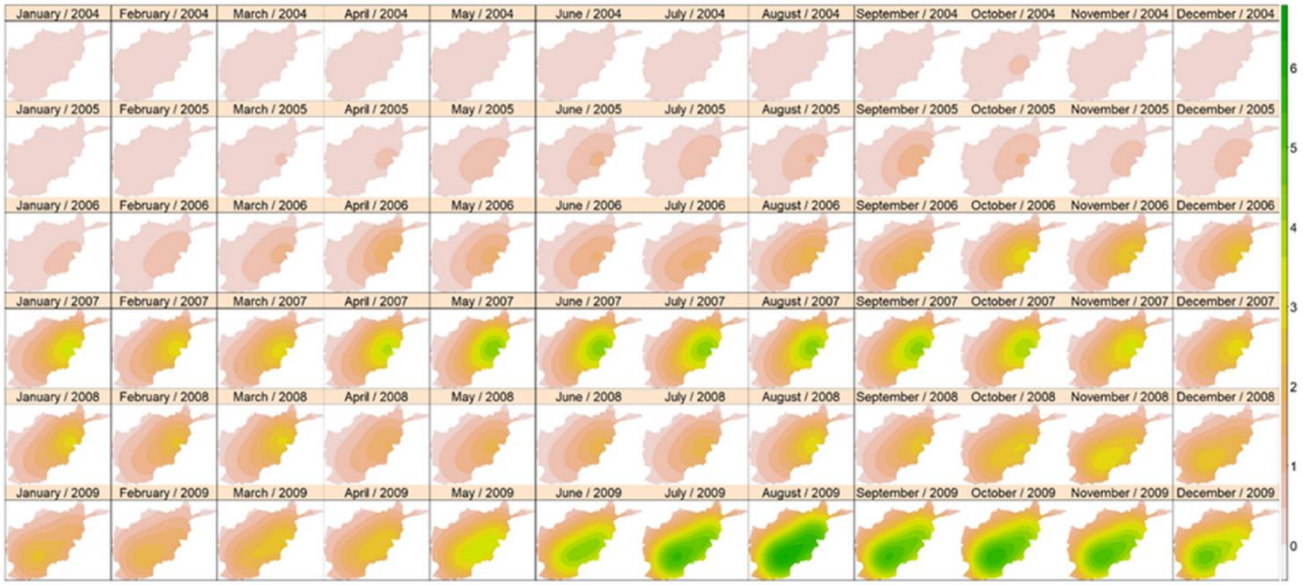

Figura 4.1. Mapa de la función de densidad con la media aritmética según CvL

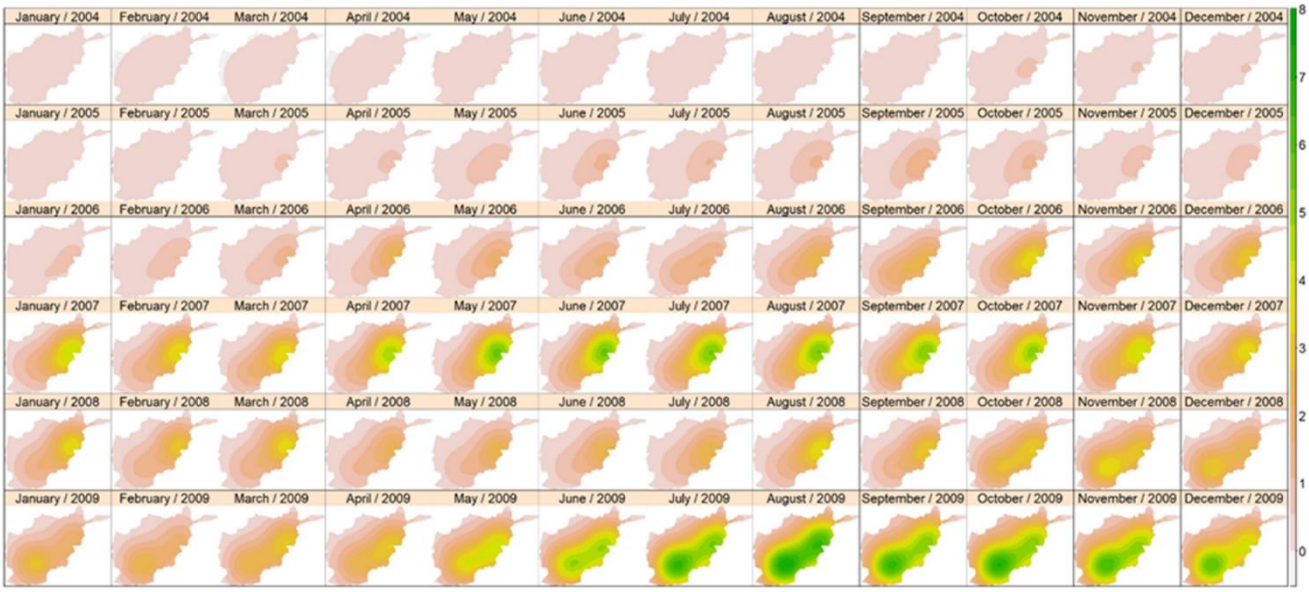

Figura 4.2. Mapa de la función de densidad con la media geométrica según CvL

Comparando ambas imágenes, (sobre todo en los meses a partir de agosto), se distinguirán más diferencias.

- *Función K:*

Cualquier función K tendrá una zona sombreada, una línea continua y una línea punteada. En el caso de R, (el programa utilizado para desarrollar el estudio) la línea continua será la K observada, la discontinua K esperada y la zona sombreada corresponderá al patrón de dispersión o al patrón de clústering en función de dónde se encuentre.

En resumen:

▪ $Kobs$ por encima de la banda gris = clúster

▪ $Kobs$ por debajo de la banda gris = distribución dispersa

▪ $Kobs$ dentro de la banda gris = Proceso Poisson

Para demostrar que los puntos de este data set tienden a formar clústers y poder implementar correctamente el método elegido, se han obtenido 72 funciones K diferentes, una para cada subconjunto (para cada mes de cada año) y una para cada criterio; es decir, que se han obtenido 72 funciones K para los subconjuntos bajo el criterio de Scott, 72 funciones K bajo el criterio de CvL y 72 funciones K bajo el criterio de validación cruzada. Todas estas representaciones de K pueden consultarse en el punto 6 de este mismo documento, excepto las funciones K obtenidas bajo el criterio de validación cruzada, ya que no aportan información relevante y se han descartado.

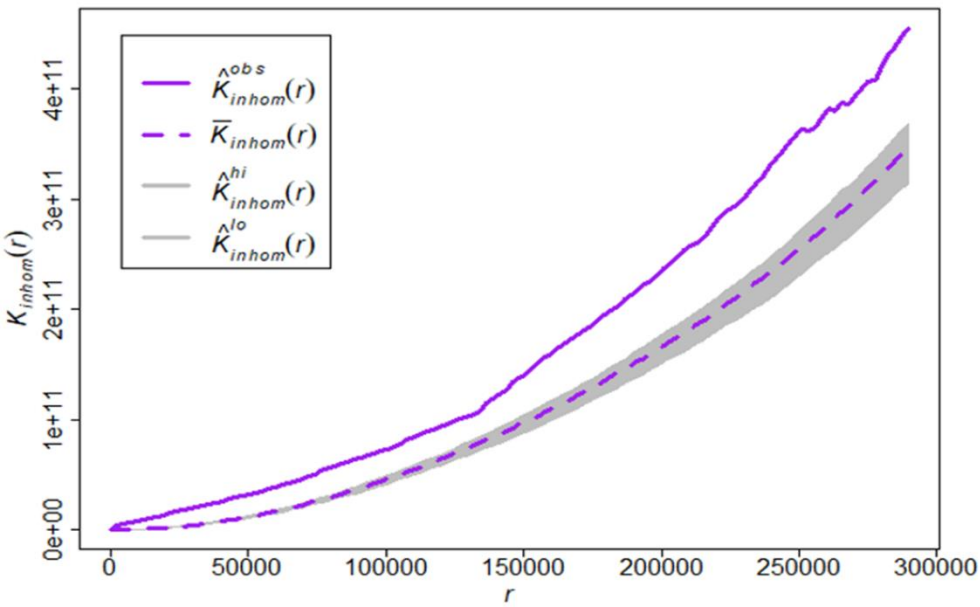

Figura 4.3. Representación de la función K para el subconjunto de datos 54 según el criterio de Scott

5. *Ideas generales:*

Los procesos puntuales son una parte de la estadística espacial que consiste en un proceso estocástico conformado de puntos en el plano; es decir, **son un conjunto de datos que se encuentran en una región concreta.** Cuando x = {x1, x2, ..., xn}, $0 \leq n \leq \infty$, un patrón puntual observado de un proceso puntual simple (es decir, que no contiene varios eventos por localización) y siendo X finito en R2 ; para cualquier conjunto escogido al azar $A \subset R2$, el cardinal de X lo proporciona la función de conteo:

$$N(X \cap A) = \sum_{x \in X} \mathbf{1}\{x \in A\} < \infty.$$

Por otro lado, la fórmula de Campbell establece que para cualquier función medible $f : R2 \rightarrow [0, \infty)$, se cumple lo siguiente:

$$\mathbb{E}\left[\sum_{x \in X} f(x)\right] = \int_{\mathbb{R}^2} f(u)\lambda(u)\mathrm{d}u,$$

Donde λ se define como la función de intensidad de X, la cual dirige su distribución espacial.

$$\mathbb{E}\left[N(X \cap A)\right] = \int_A \lambda(u)\mathrm{d}u.$$

Si la función de intensidad es constante, el proceso dado por X será **homogéneo,** y **si no es constante,** será **inhomogéneo,** en cuyo caso la distribución espacial variará a lo largo de la función soporte. Se observa una única realización, por lo que se debe tener una estimación de la función de intensidad que pueda imitar la distribución espacial del proceso subyacente. Por esta razón se consideran distintos tipos de estimadores no paramétricos de la intensidad.

Para cada coord. cartesiana ‘x’ e ‘y’, donde las ‘s’ son las desviaciones típicas de las coordenadas bidimensionales de los eventos.

$$(s_x n^{-1/6}, s_y n^{-1/6}),$$

En este caso, se comprueba que el parámetro suavizado es un vector de dos componentes para suavizar las dos coordenadas cartesianas bidimensionales. Pero por otro lado, Cronie and Van Lieshout propusieron encontrar el parámetro óptimo minimizando la función de parámetro suavizado.

$$CvL(\sigma) = \left(|W| - \sum_{i=1}^{n} 1/\widehat{\lambda}^{*}_{\sigma}(x_i)\right)^2,$$

- *Tipo de datos usados:*

Como ya se ha mencionado anteriormente, en esta base de datos existen 72 subconjuntos distintos de datos que se extienden desde enero de 2004 hasta diciembre de 2009. Al tratarse de procesos puntuales con coordenadas, no se obtendrán tablas de valores ya que no resultaría útil para trabajar en este proyecto. Lo que sí se van a obtener son representaciones de funciones y mapas, para el correcto análisis de los datos.

- *Preparación de los datos:*

La preparación de los datos se divide en los siguientes apartados:

1. Paquetes de RStudio utilizados: Tan sólo será necesario cargar la librería 'Spatstat', ya que al cargarla, el mismo programa se encarga de ejecutar todos los paquetes relacionados que se van a usar posteriormente.

2. Carga del data set: Al tratarse de un archivo con extensión .pp, no es necesario cargar ningún paquete adicional, ni programar una ruta específica para cargar el data set dentro de RStudio. Basta con cargar el data set directamente desde la consola.

3. Creación de los bandwidth: Durante todo el proyecto será necesario apoyarse en los bandwidth, por lo que habrá que crear tres distintos con la función 'lapply': uno para Scott, uno para validación cruzada y otro para CvL. Cada bandwidth se guarda como una variable independiente correctamente identificada. Los nombres de los bandwidth serán: bw_CvL, bw_ppl y bw_Scott.

4. Obtención de la media aritmética y geométrica: La media aritmética se obtiene gracias a la función 'median', mientras que la media geométrica es calculada con la función 'exp(mean(log))'.

5. Funciones de densidad: Las funciones de densidad, tanto para la media aritmética como para la media geométrica se han obtenido con la función 'density.ppp', donde en sigma se ejecuta el bandwidth, leaveoneout = FALSE, diggle = TRUE, positive = TRUE.

6. Ráster: La creación de los ráster es el último paso para poder plotear los mapas de Afganistán correctamente. Se crea un ráster por cada criterio, es decir, uno para Scott, otro para ppl y otro para CvL.

7. Funciones K: Las funciones K se calculan mediante la función de RStudio 'envelope'. Cabe destacar que será necesario implementar un código para crear la función K de los 72 subconjuntos: de CvL calculado con la media geométrica, de Scott calculado con la media geométrica, y así mismo de ppl. En total, el trabajo debe de contener 216 funciones K.

6. *Modelización:*

En primer lugar, se calculan las funciones de densidad usando la media aritmética bajo el criterio de Scott, se crea un ráster y se plotean los mapas en función de los meses, obteniéndose la siguiente figura:

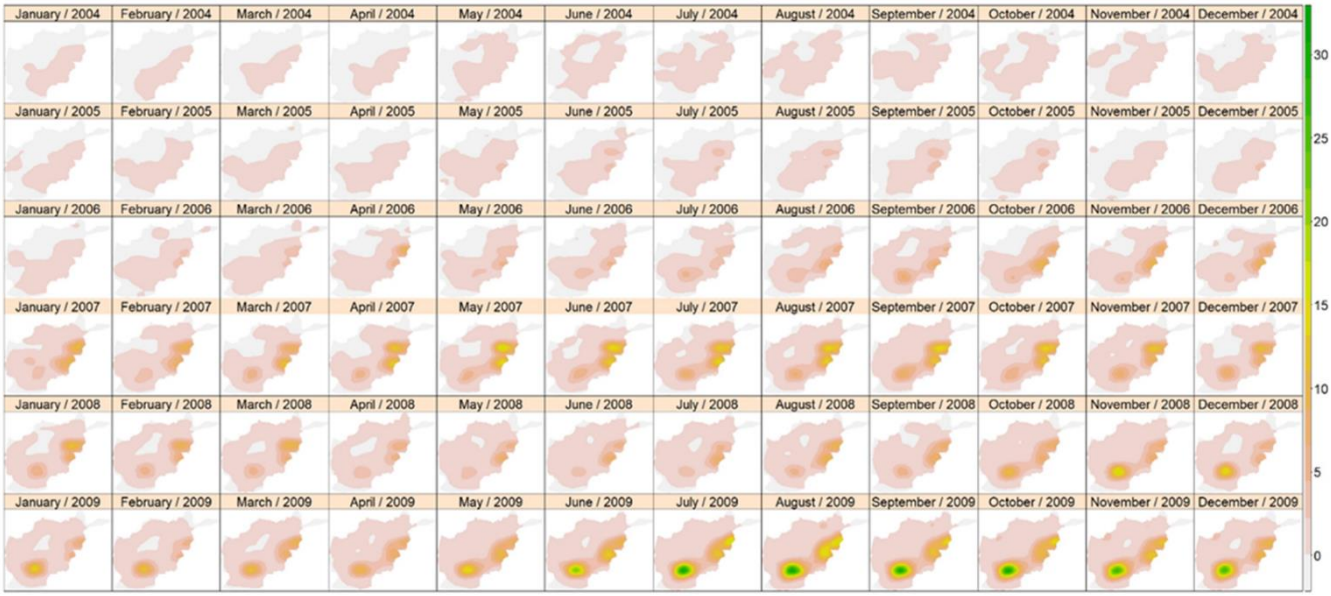

Figura 6.1. Representación de la función de densidad para el data set según el criterio de Scott, usando la media aritmética.

Se puede observar con claridad cómo escala el conflicto según van pasando los meses, aunque este modelo no es el definitivo, pero ya establece un patrón de comportamiento de los datos claramente definido.

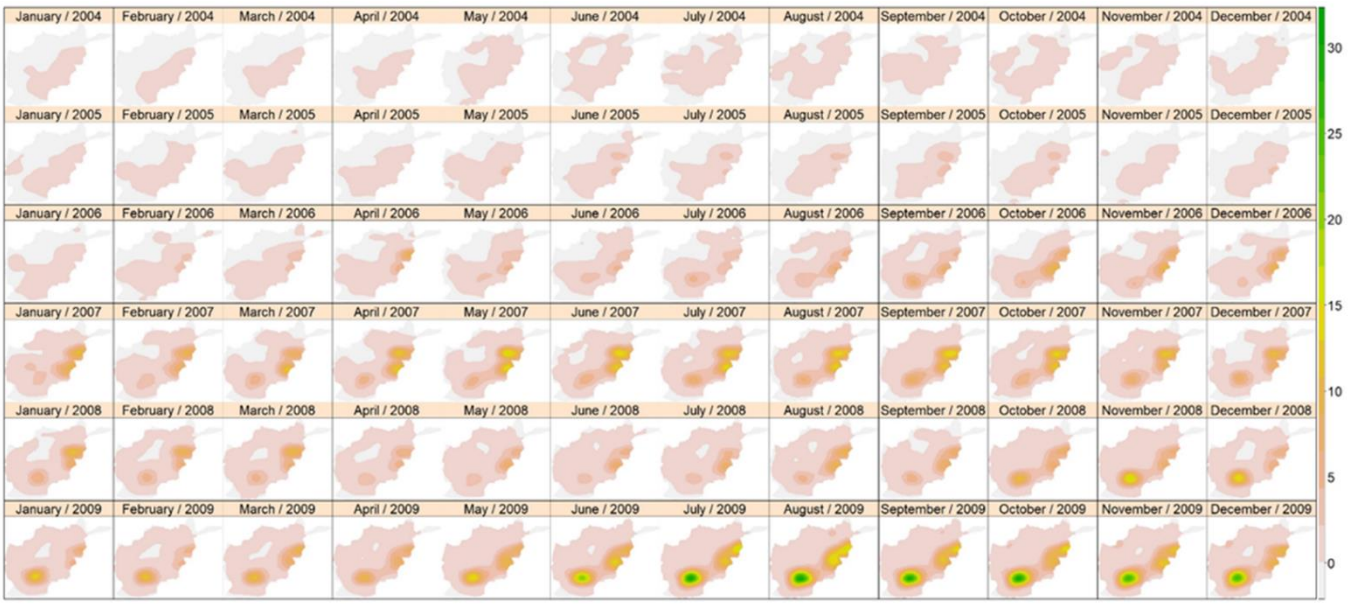

Figura 6.2. Representación de la función de densidad para el data set según el criterio de Scott, usando la media geométrica.

Aunque a priori ambas figuras parezcan idénticas, existen algunas diferencias, pero no demasiado significativas. Por ejemplo, para el mes de octubre de 2008, cuando se calcula la función de densidad se observa una pequeña mancha blanca en la imagen correspondiente a la media aritmética, mientras que para la media geométrica no se observa. Esto es consecuencia de que ambas medias, aunque difieren, son muy próximas entre sí.

En cambio, para las figuras obtenidas bajo el criterio de CvL se observan algunos matices que bajo el criterio de Scott no se observan. Estas son las figuras obtenidas:

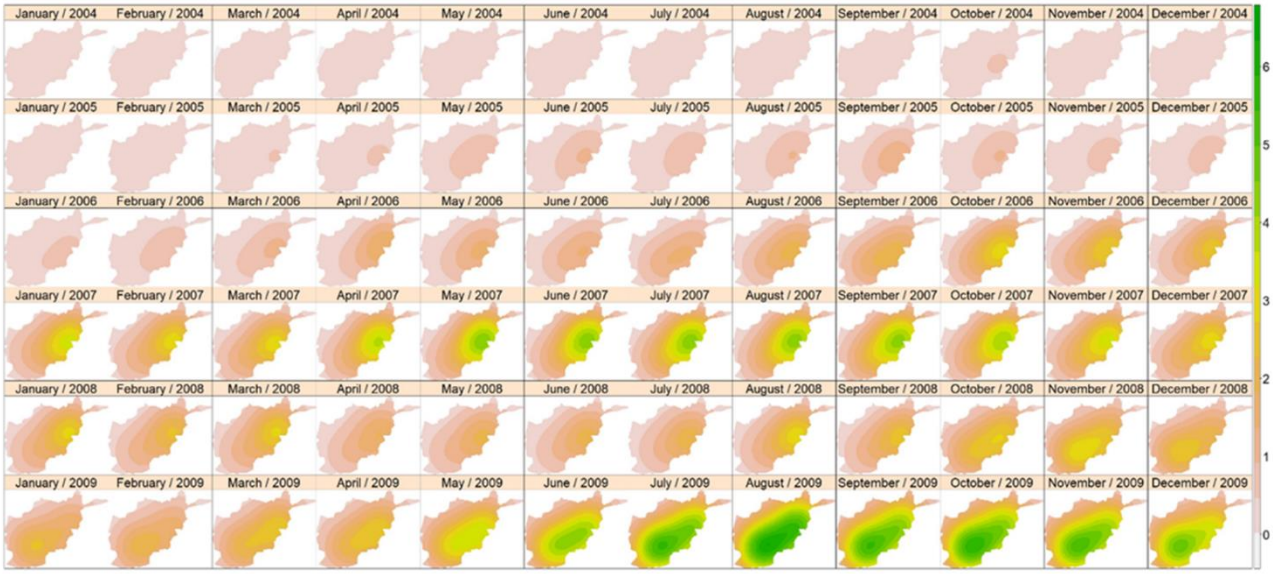

Figura 6.3. Representación de la función de densidad para el data set según el criterio de CvL, usando la media aritmética.

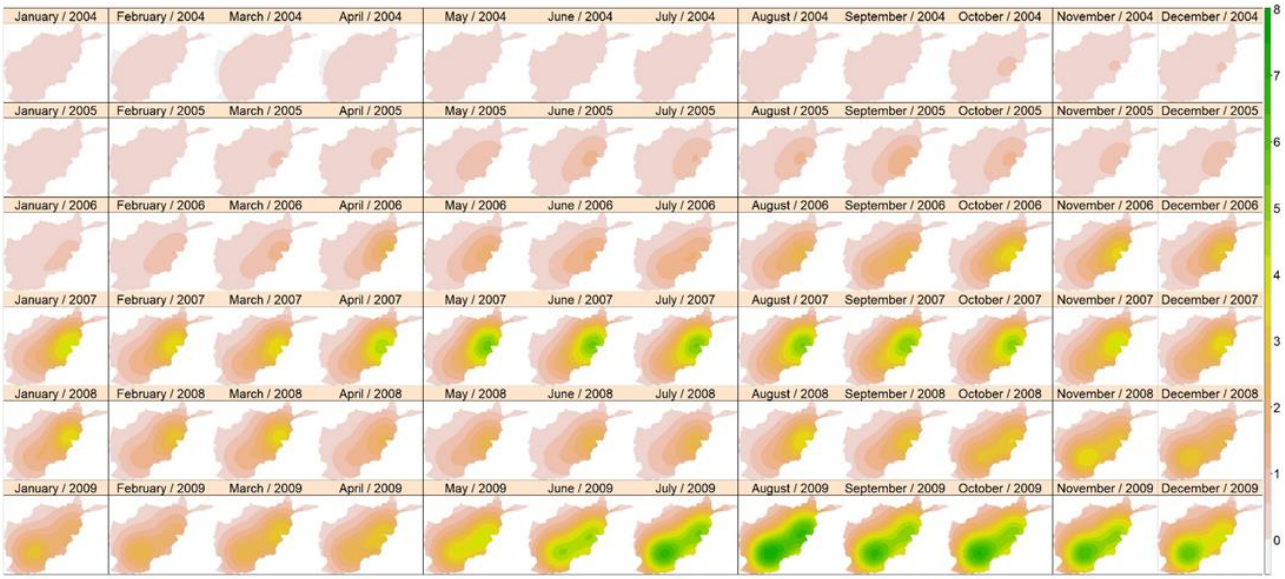

Figura 6.4. Representación de la función de densidad para el data set según el criterio de CvL, usando la media geométrica.

A partir de Junio de 2009, si se comparan ambos mapas, resaltan algunas diferencias que bajo la metodología de Scott no se detectaban. Por ejemplo, en el caso del mapa obtenido con la media geométrica pueden observarse dos clústers verdes mejor diferenciados respecto al mismo mes en el mapa obtenido con la media aritmética. **Esto, a priori, podría interpretarse como una mejor obtención del punto concreto donde se produce la escalada del conflicto.**

Finalmente, también se han obtenido estas representaciones bajo el criterio de validación cruzada. Corresponden a las siguientes figuras:

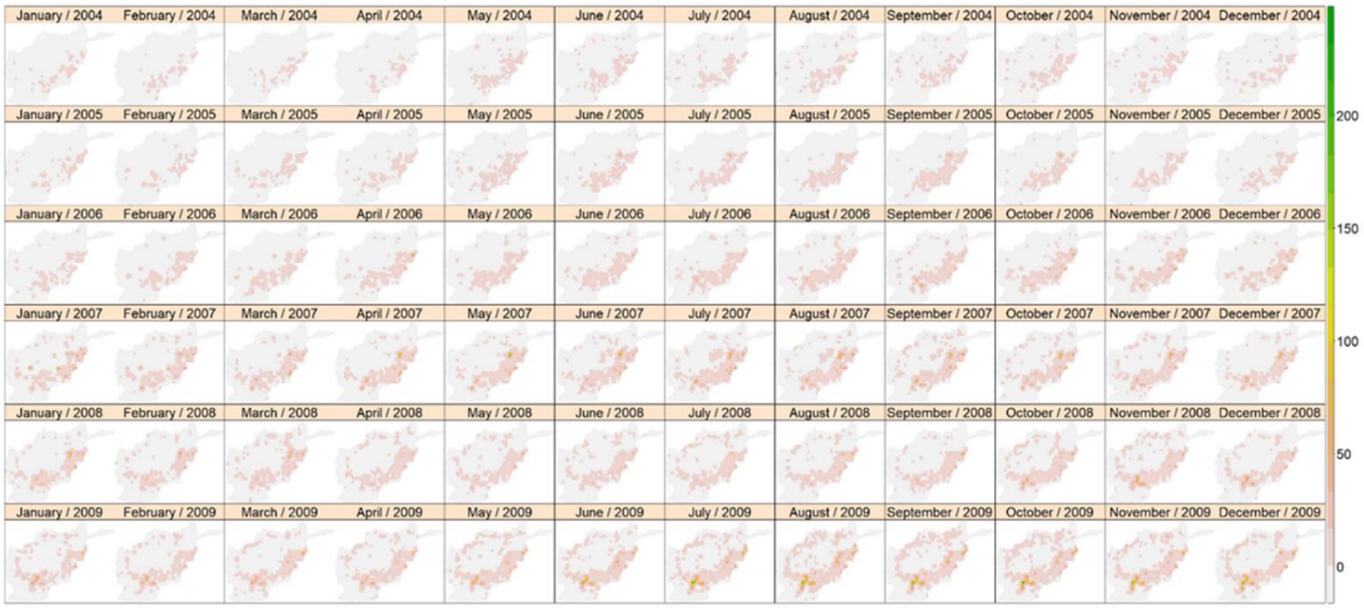

Figura 6.5. Representación de la función de densidad para el data set según el criterio de ppl, usando la media aritmética.

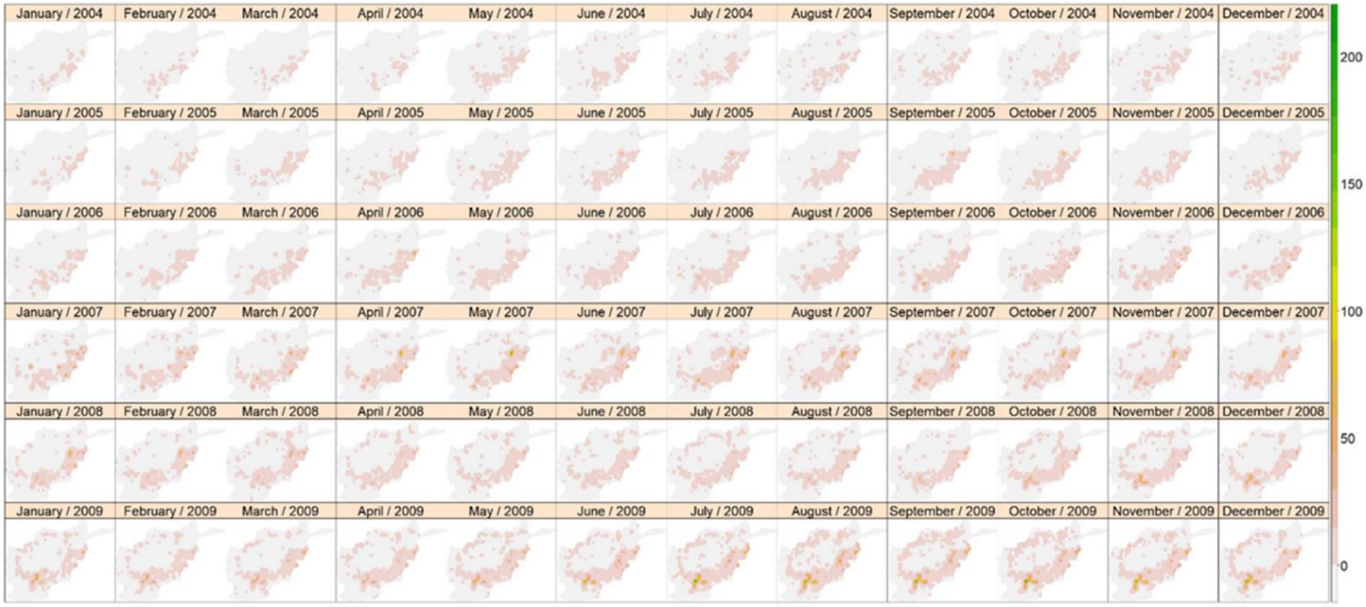

Figura 6.6. Representación de la función de densidad para el data set según el criterio de ppl, usando la media geométrica

No nos da apenas información; por esta razón no se han incluido las funciones K para el método de validación cruzada.

7. Evaluación

Como ya se ha definido anteriormente, la función K define que **si K_{obs} cae encima de la banda gris, la distribución será más agrupada** que una distribución aleatoria, mientras que **si K_{obs} cae por debajo, la distribución será más dispersa** que una distribución aleatoria en esa distancia y no se podrán formar clústers. Esto significa que **si el resultado obtenido es una distribución dispersa, los datos no tenderán a formar clústers** y por tanto, no se podrá saber en qué momento ni en qué lugar se produce la escalada.

Veamos un ejemplo de función K obtenida para el punto álgido de la escalada del conflicto: agosto de 2009, bajo el criterio de Scott:

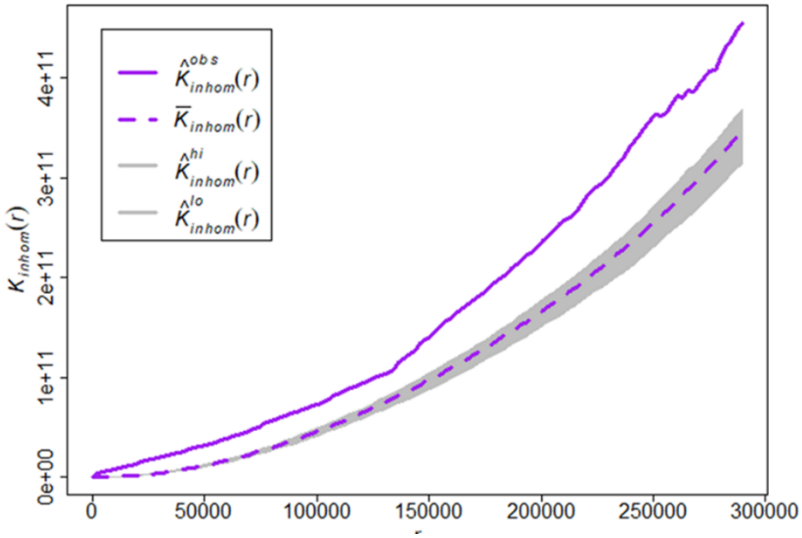

Figura 7.1. Representación de la función K para el subconjunto de datos 68 (agosto de 2009) según el criterio de Scott

Como Kobs está por encima de la banda, la distribución será más agrupada que una distribución aleatoria, lo que implica la formación de clústers representando eventos de la escalada.

Podemos observar que para CvL ocurre exactamente lo mismo:

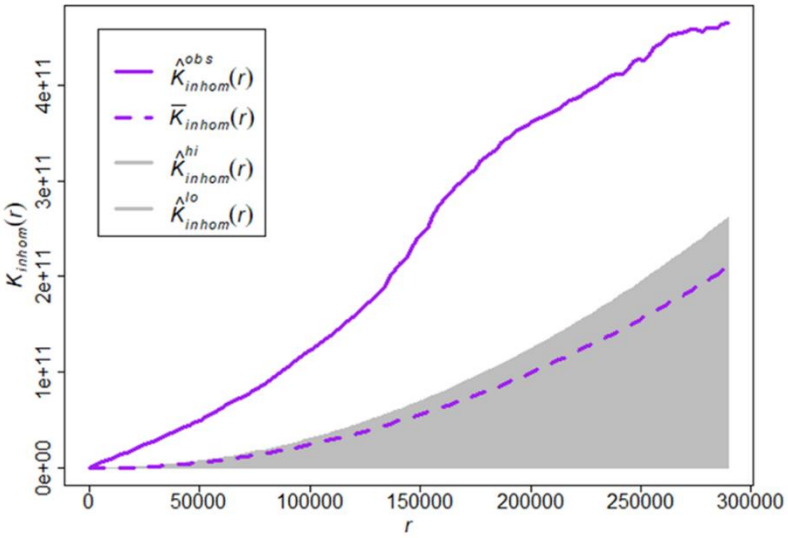

Figura 7.2. Representación de la función K para el subconjunto de datos 68 (agosto de 2009) según el criterio de CvL

En las siguientes representaciones obtenidas se puede comprobar cuál es la evolución de la función K para los criterios de Scott y CvL, para poder extraer las conclusiones oportunas.

Como el objetivo de este proyecto es hallar el momento y lugar exactos donde se produce la escalada del conflicto de Afganistán, lo más coherente es comparar las funciones K obtenidas por año y estudiar su evolución en función del avance de los meses. Así mismo, en caso de aparecer un patrón "extraño" respecto a los patrones considerados dentro de la normalidad, se detectará y se evaluará su comportamiento para determinar si interfiere con los datos.

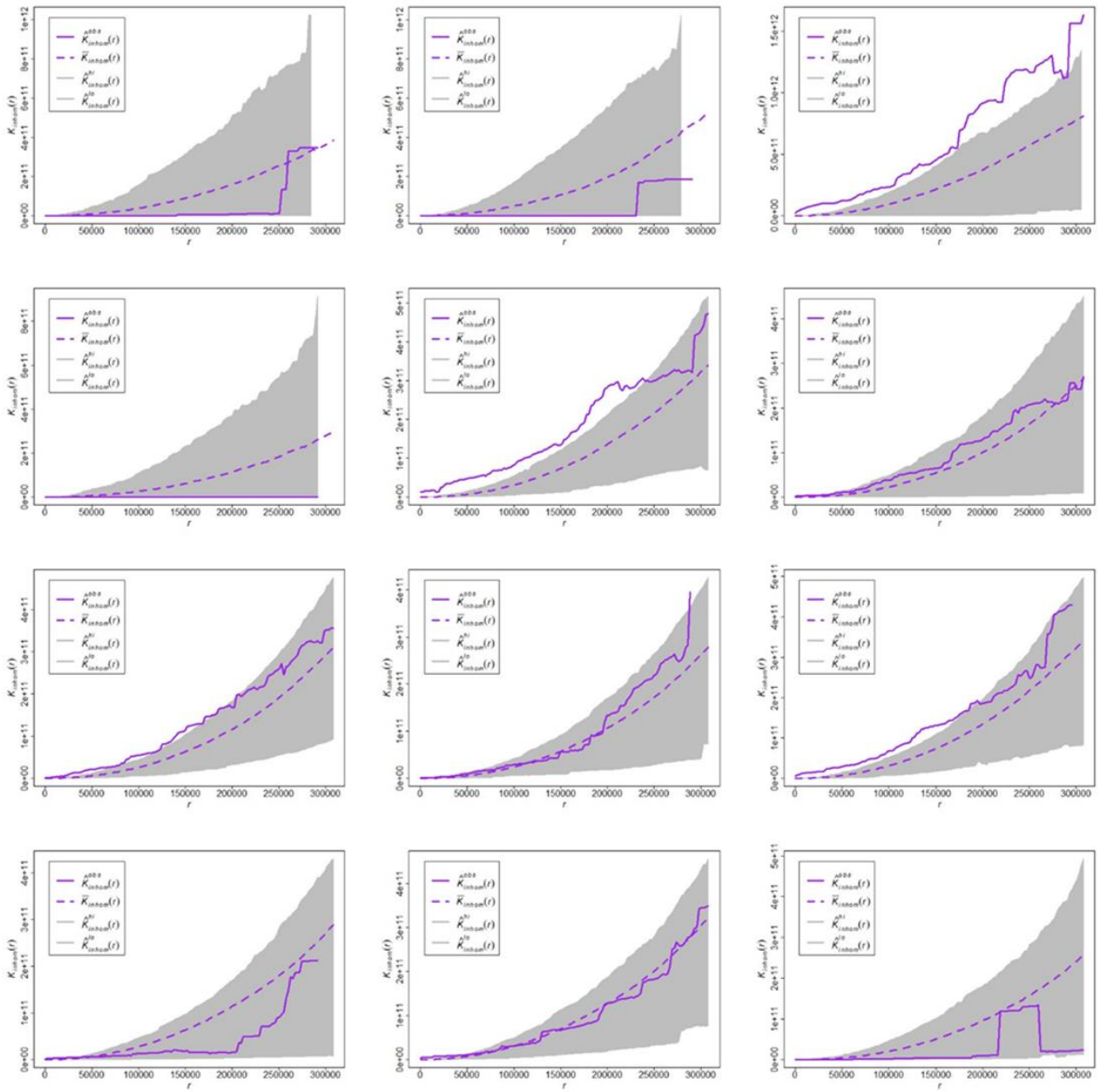

Figura 7.3. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2004 según el criterio de Scott

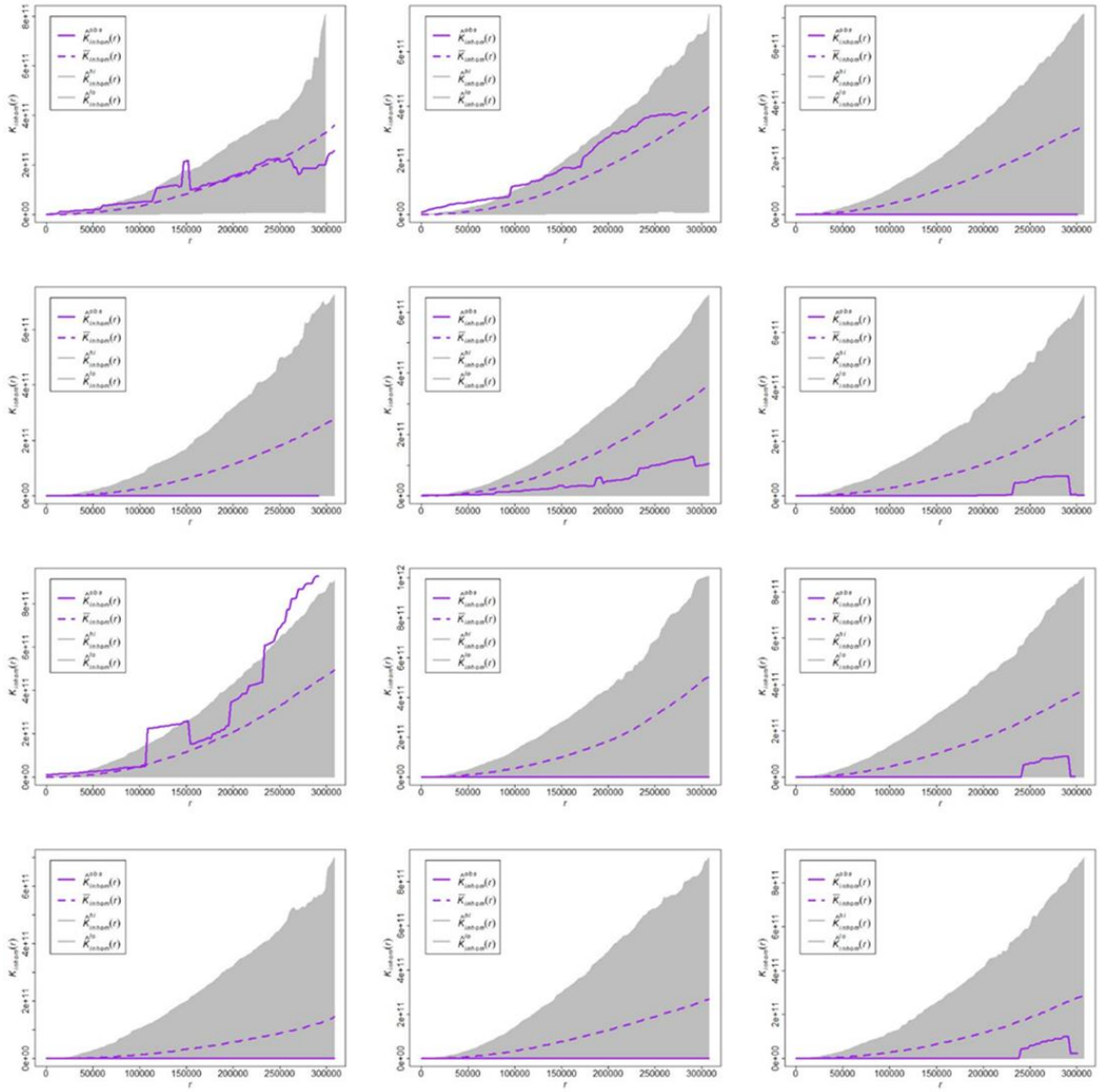

Figura 7.4. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2005 según el criterio de Scott

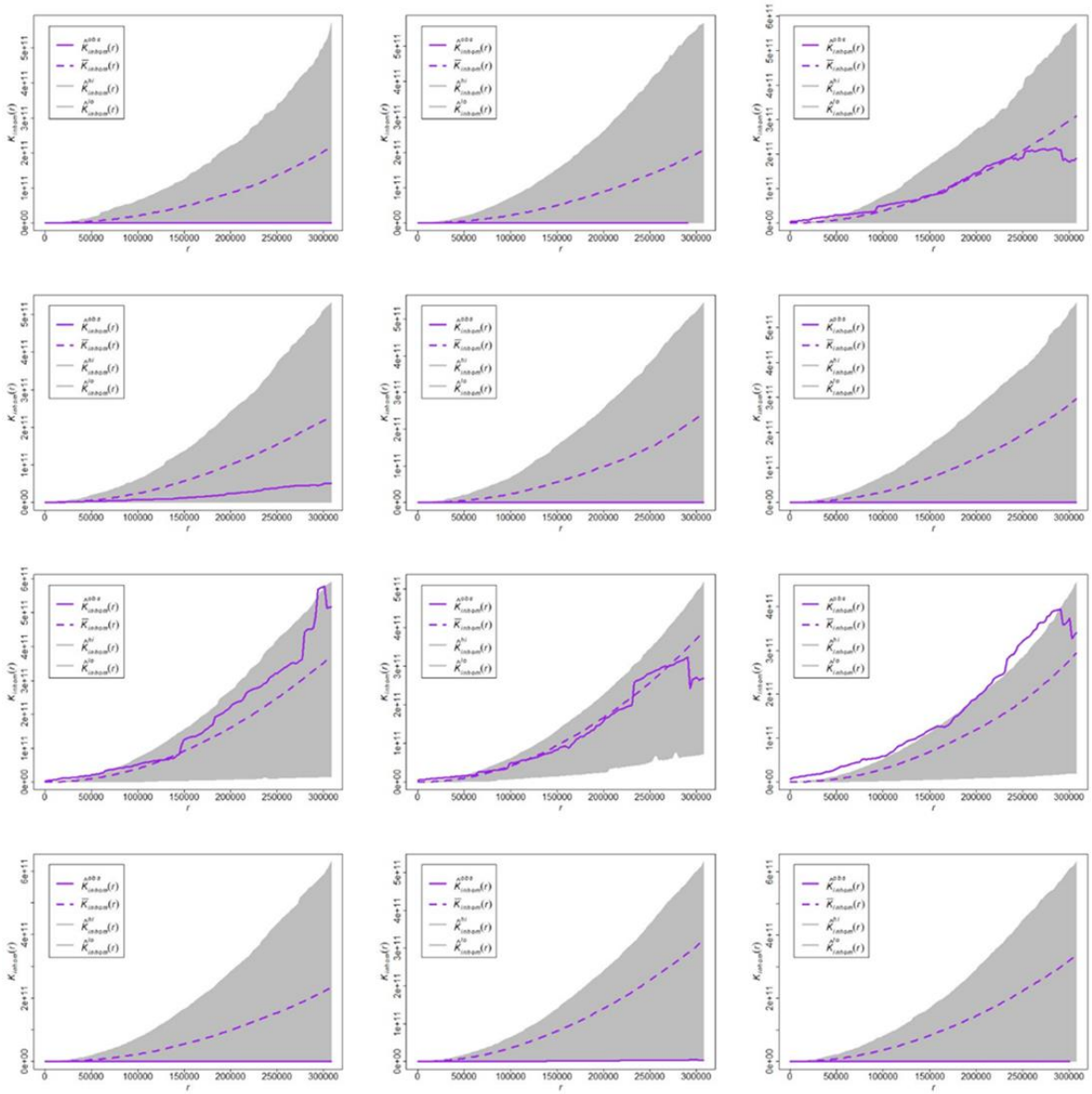

Figura 7.5. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2006 según el criterio de Scott

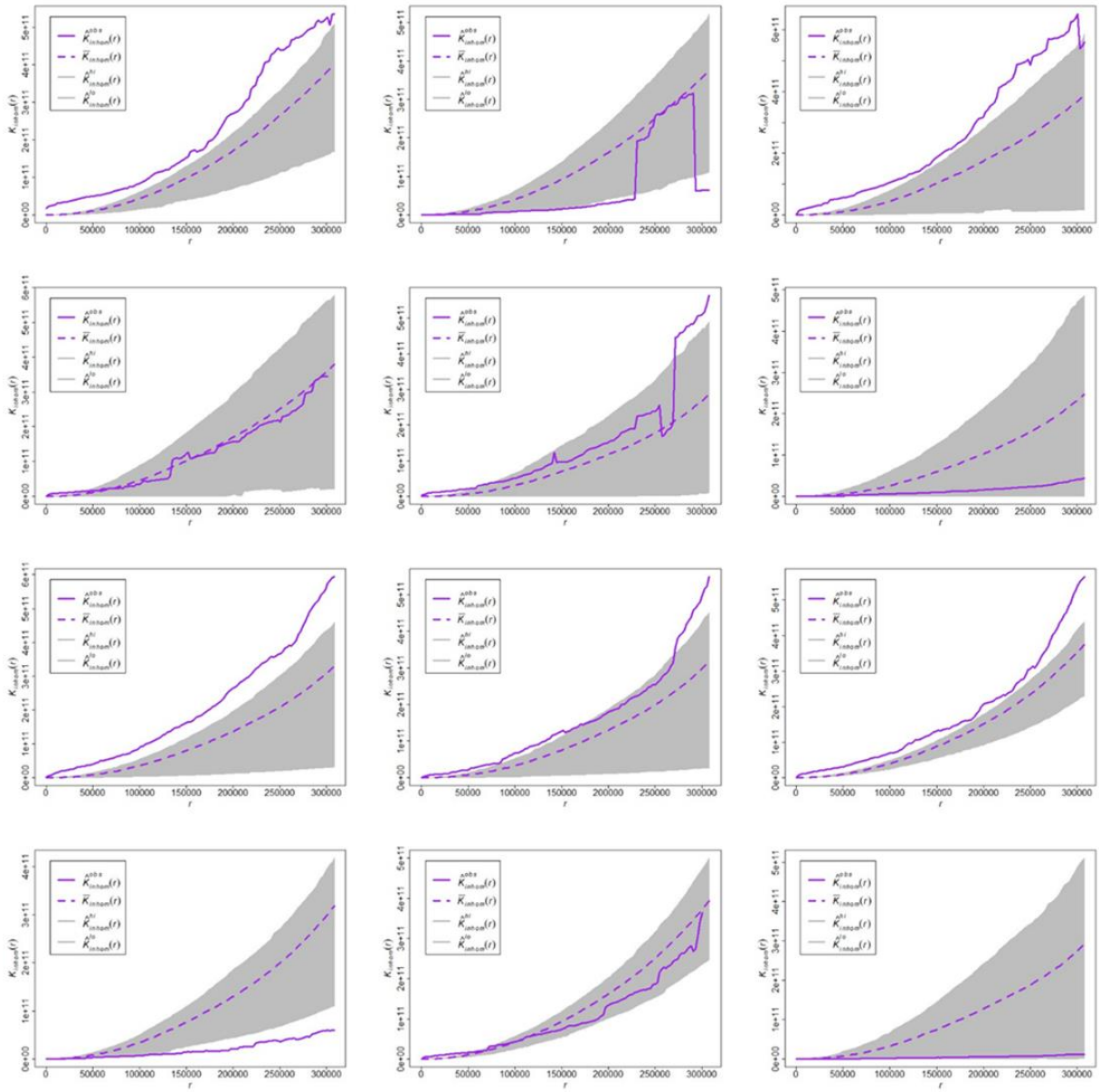

Figura 7.6. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2007 según el criterio de Scott

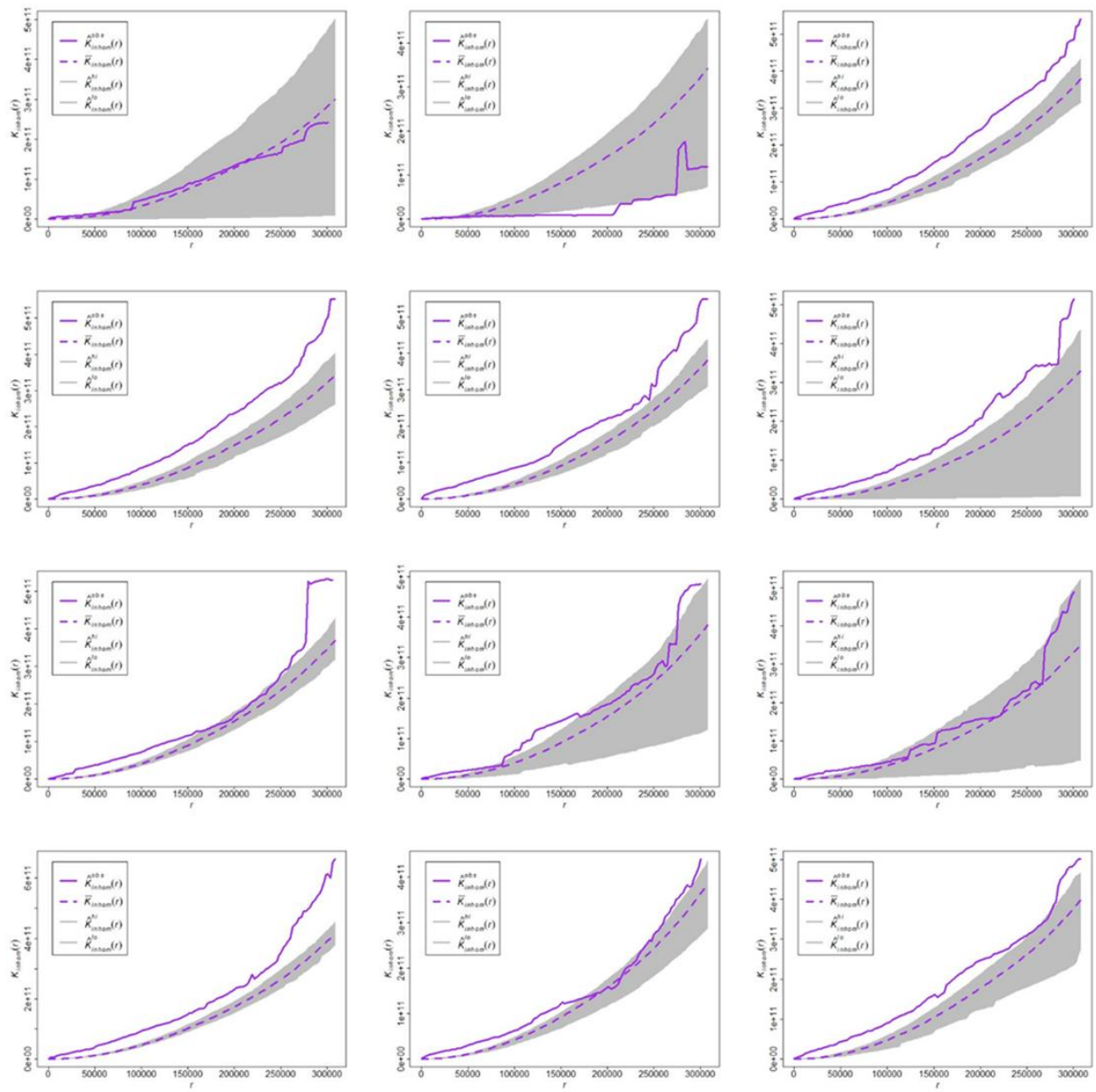

Figura 7.7. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2008 según el criterio de Scott

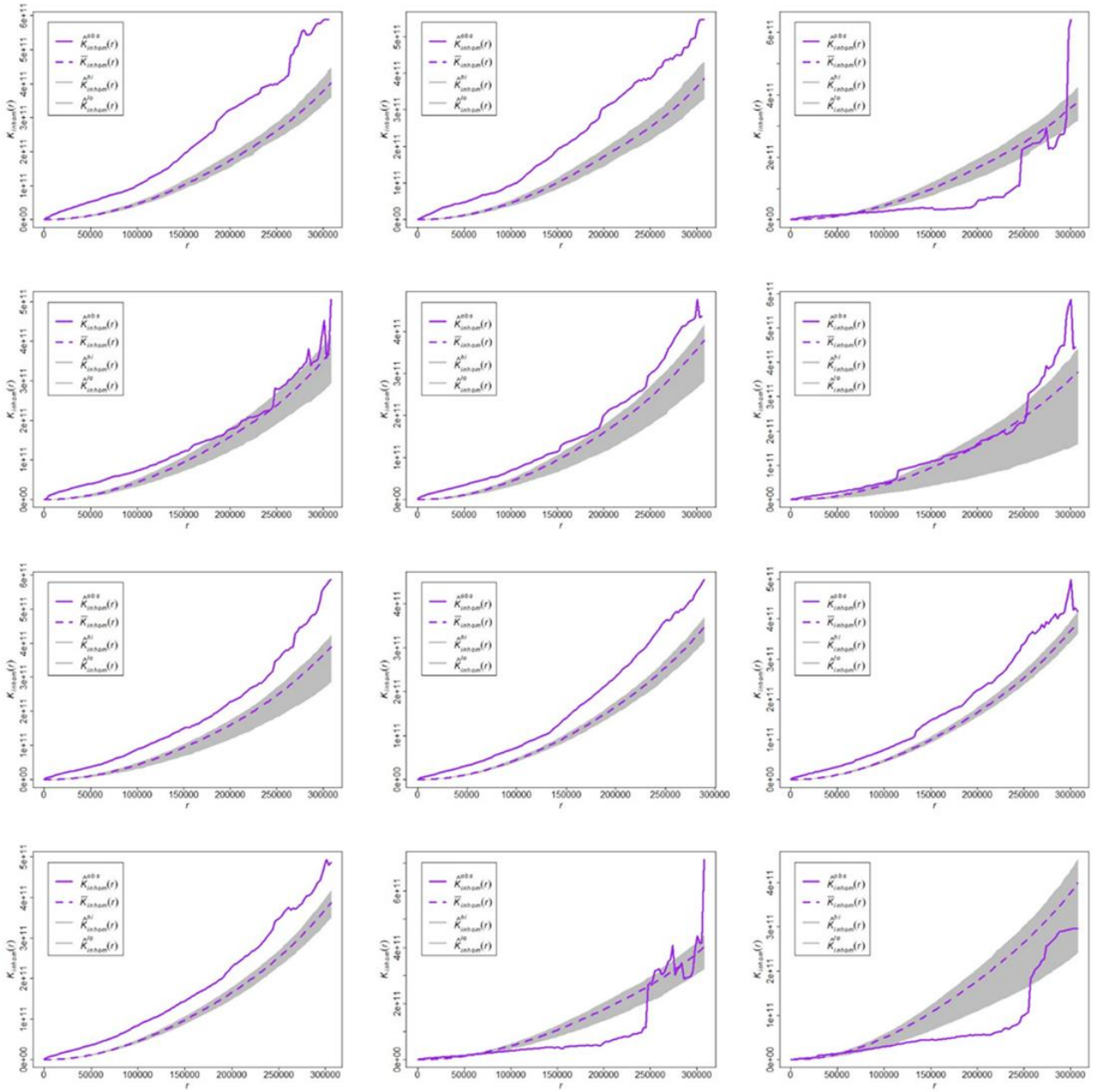

Figura 7.8. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2009 según el criterio de Scott

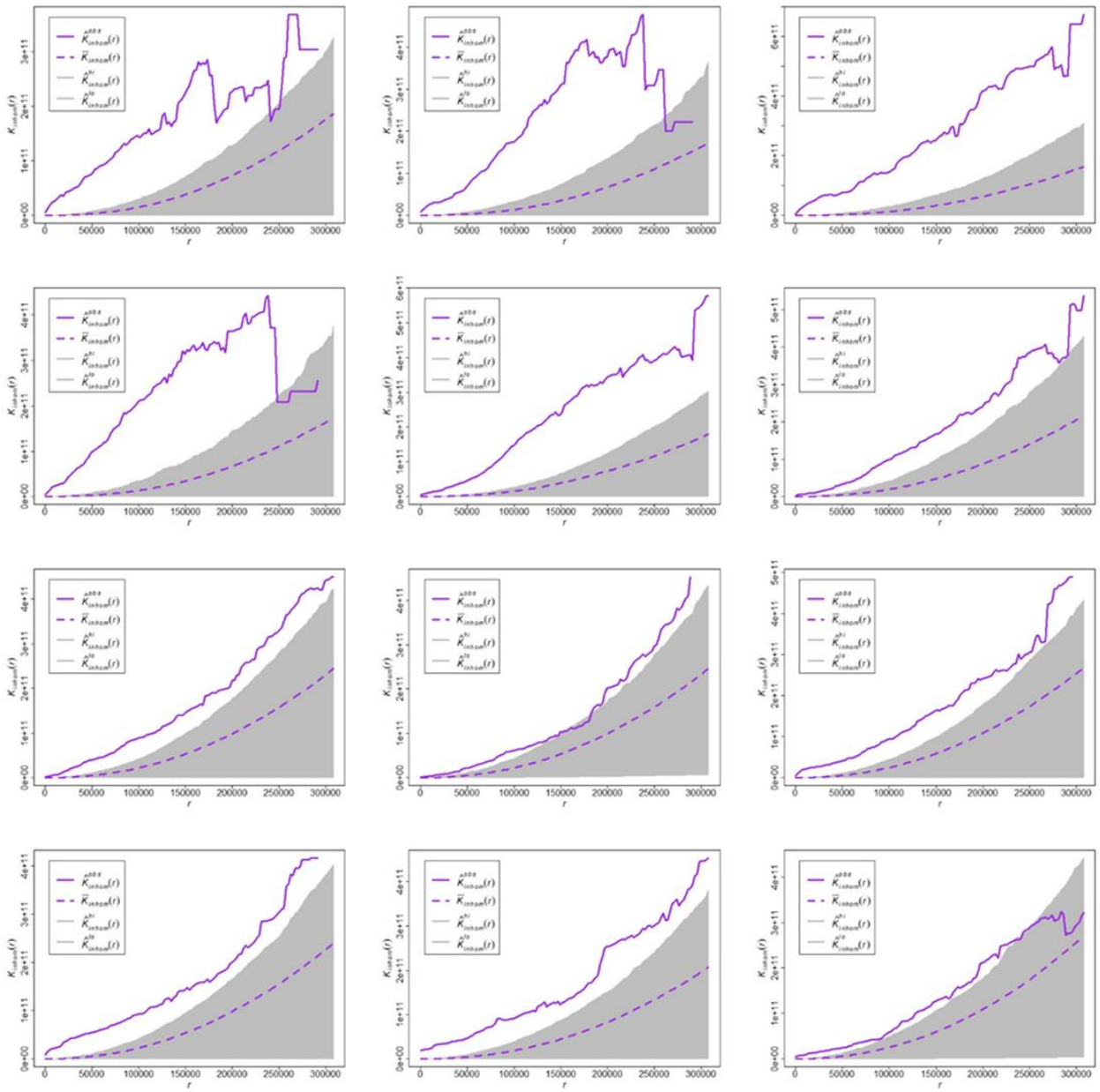

Figura 7.9. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2004 según el criterio de CvL

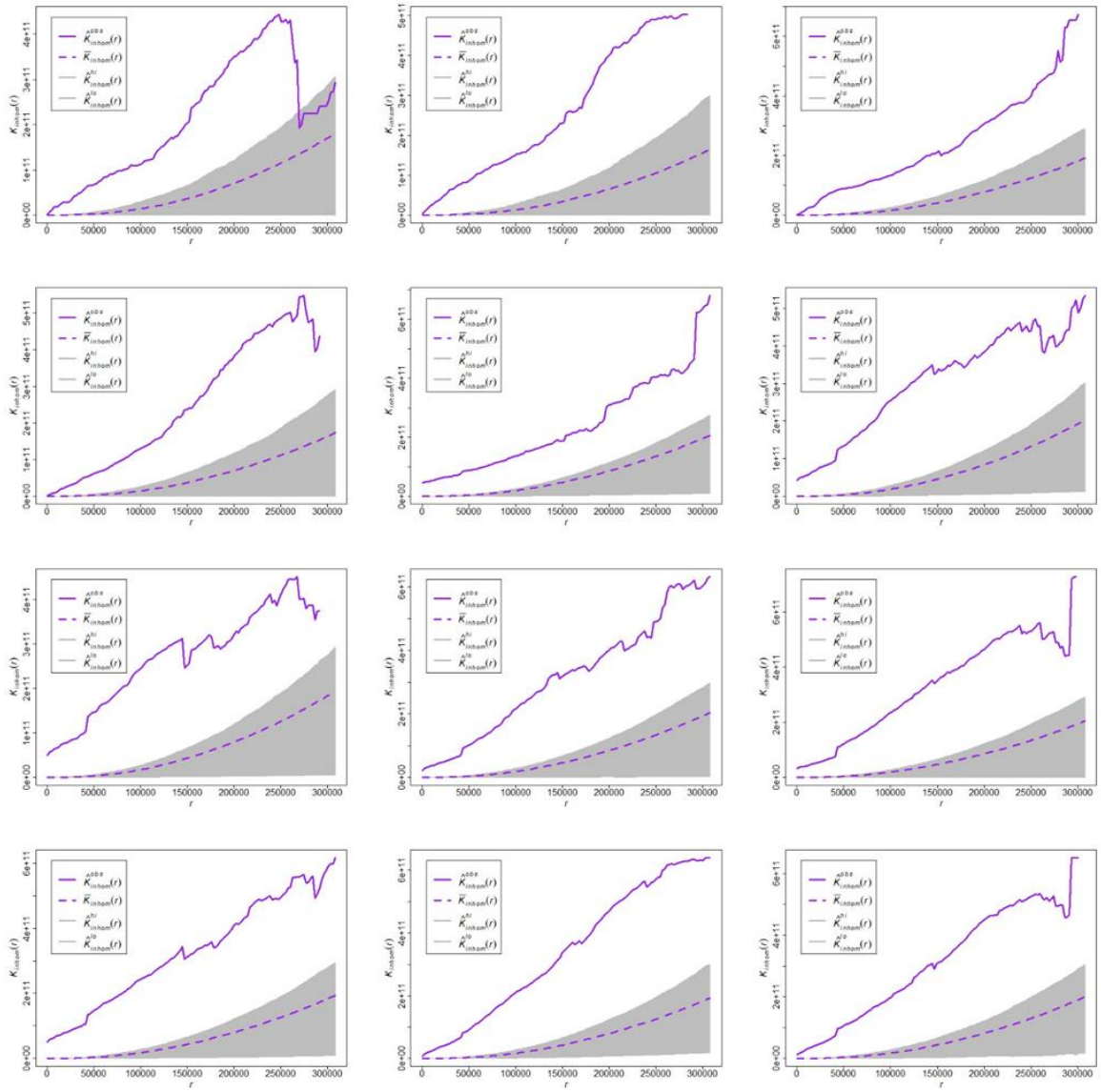

Figura 7.10. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2005 según el criterio de CvL

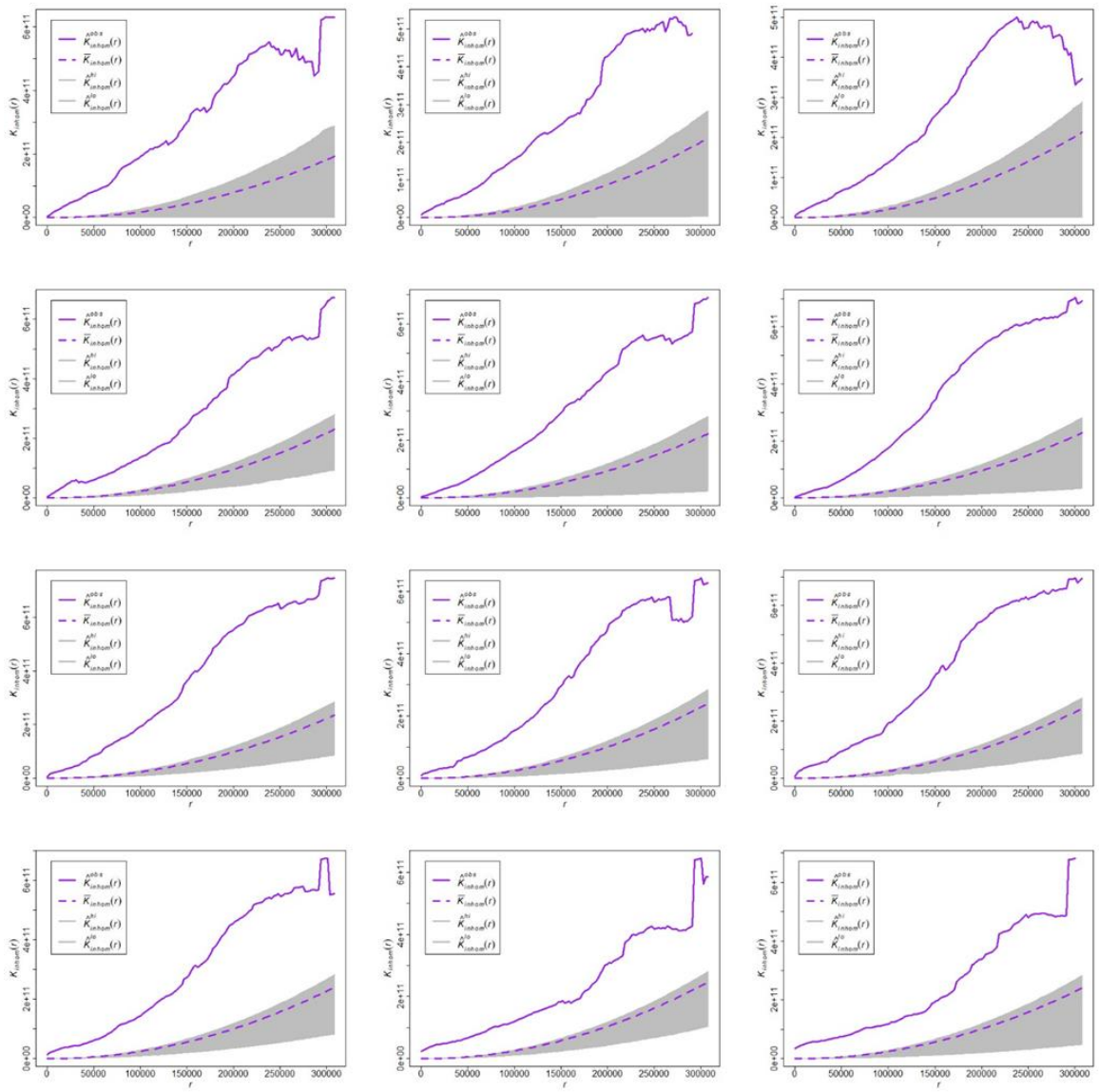

Figura 7.11. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2006 según el criterio de CvL

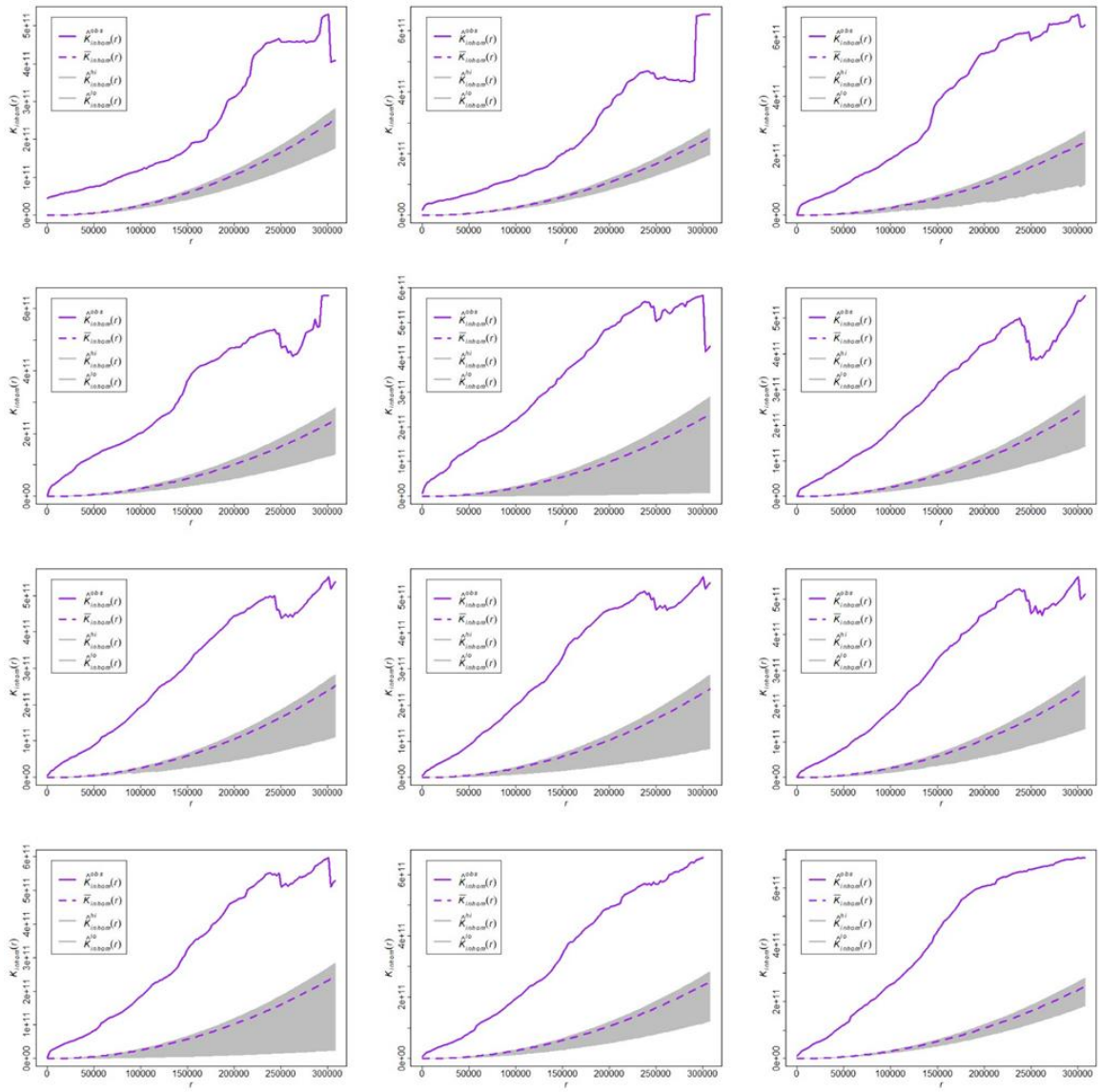

Figura 7.12. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2007 según el criterio de CvL

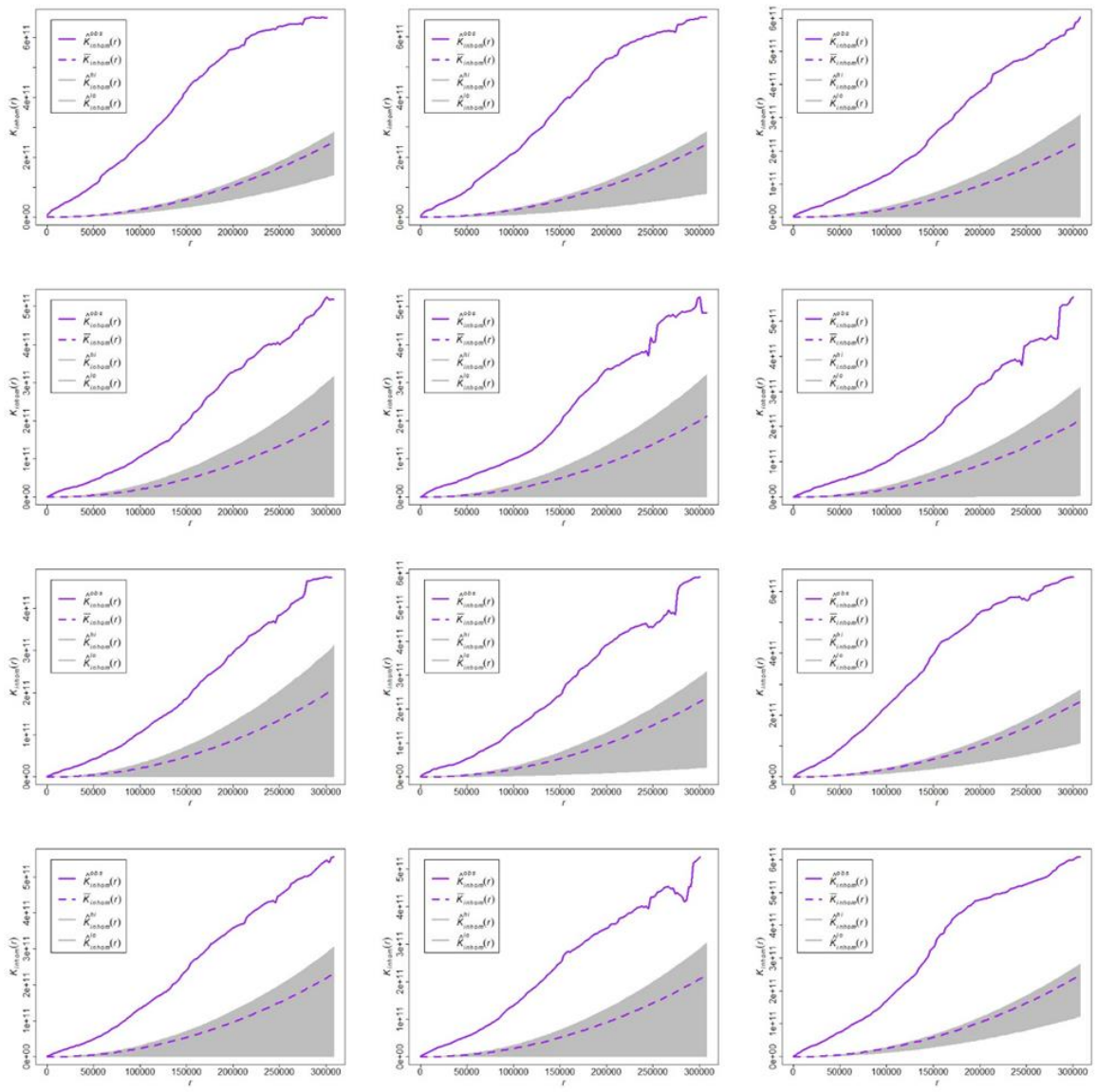

Figura 7.13. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2008 según el criterio de CvL

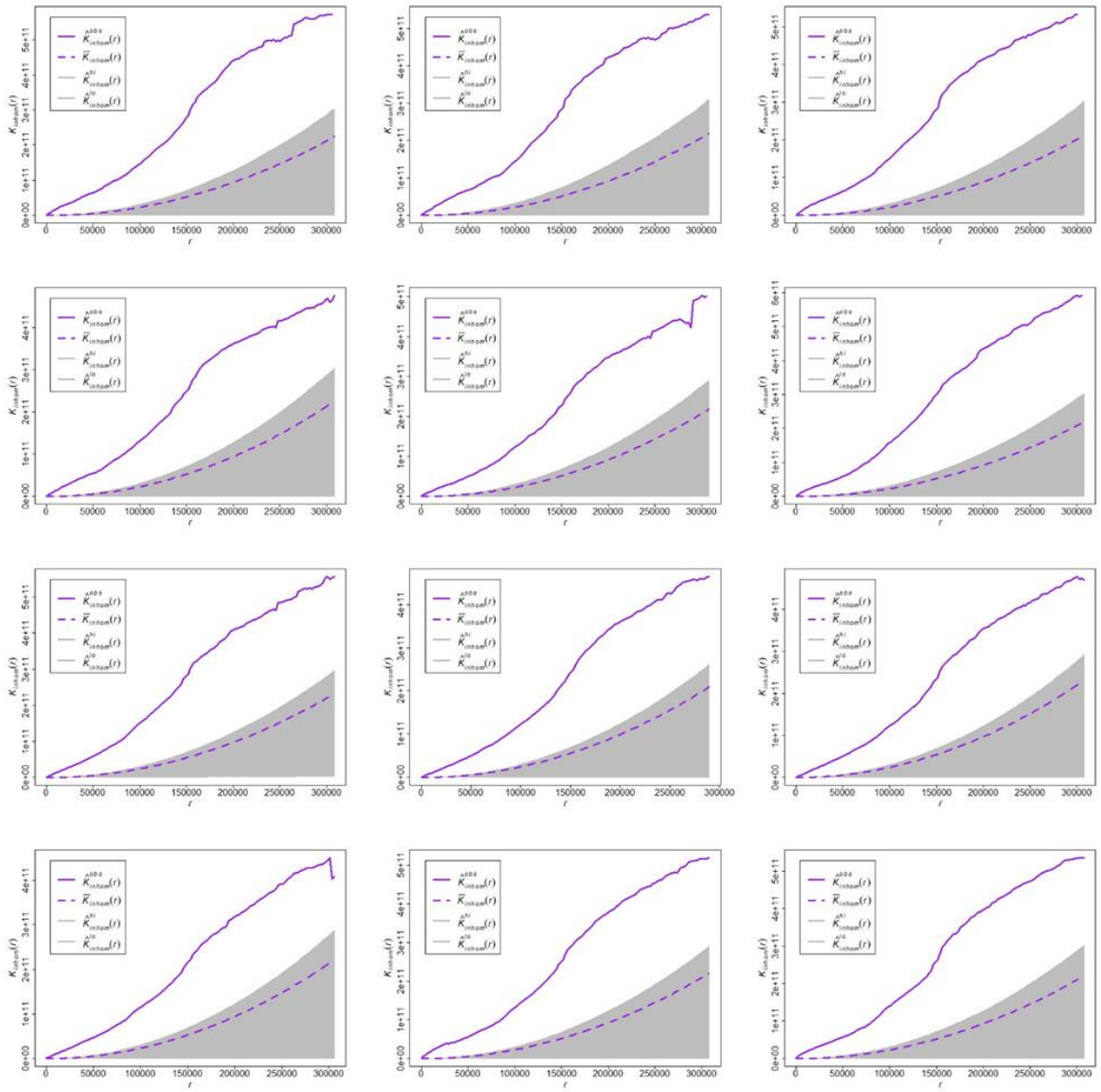

Figura 7.14. Representación de la evolución de la función K para el subconjunto de datos correspondiente a 2009 según el criterio de CvL

Como hemos podido comprobar, el método desarrollado bajo el criterio de validación cruzada es el primero en ser descartado ya que, al plotearse, no aporta información relevante sobre el estudio. No muestra una evolución de la escalada del conflicto en función del tiempo, ni las zonas concretas donde ocurre; tan sólo muestra algunos puntos aleatorios que no sirven para implementar el método. Se puede ver en el ejecutable de RScript.

No obstante, el modelo de CvL no se ajusta correctamente a los datos ya que cuando se produce la escalada del conflicto, se observa una gran mancha verde en el mapa que apenas aporta información reseñable sobre los puntos concretos donde se produce la escalada. Aun así, se observa un mejor ajuste cuando se usa la media geométrica frente a la aritmética. Puede observarse lo descrito en la siguiente figura:

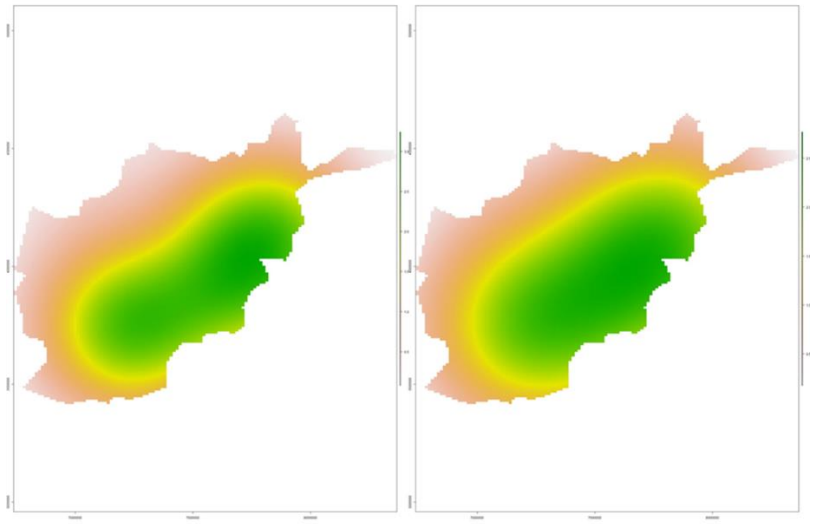

Figura 7.15. Representación comparativa de la función de densidad para el subconjunto de datos 68 (agosto de 2009) según el criterio de CvL.

Las diferencias bajo el criterio de Scott para ambas medias apenas son significativas, se puede observar en la tabla 7.1 del trabajo. Menor error cometido en los cálculos del método de Scott y un mayor ajuste del modelo a los datos.

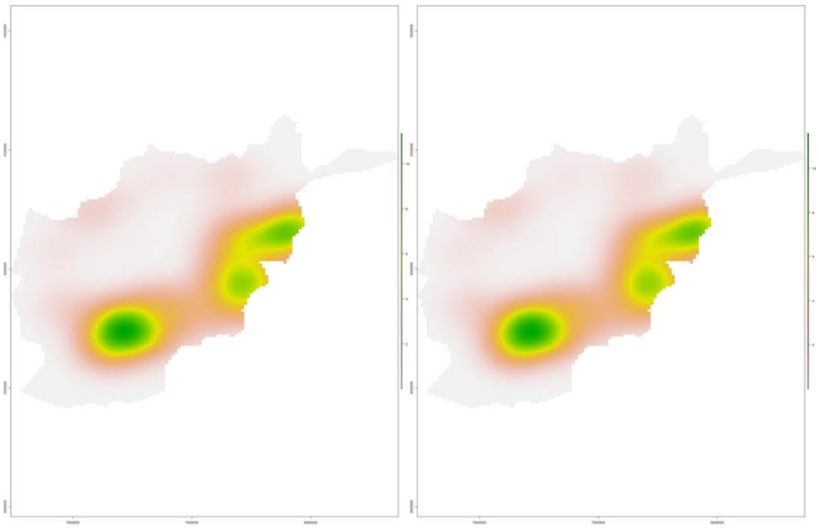

Figura 7.16. Representación comparativa de la función de densidad para el subconjunto de datos 68 (agosto de 2009) según el criterio de Scott.

	Media Aritmética	**Nombre Guardado**	**Media Geométrica**	**Nombre Guardado**
Scott	83919.17 57637.69	bw_scot_1	85404.12 58683.08	geom_bw_scot_comun
CvL	280321.1	bw_CvL_1	225041	geom_bw_CvL_comun
ppl	13789.43	bw_ppl_1	15111.34	geom_bw_ppl_comun

Tabla 7.1. Medias aritmética y geométrica según cada criterio

8. Conclusiones:

Se ha concluido que el modelo que más se ajusta a la modelización de los datos es el de Scott, ya que sus medias geométricas y aritméticas son muy cercanas entre sí, se comete menor error, y además sus puntos tienden a formar clústers. Gracias a los mapas obtenidos, puede deducirse con claridad que el mes donde se producen más conflictos y por tanto, cuando se produce la escalada, es en agosto de 2009; y la zona más afectada es el sureste del país. En los mapas de Scott se puede ver cómo hay tres puntos de mayor interés: uno al sur del país, (que es el más llamativo e intenso) y dos que se sitúan al este.

También se detecta una escalada del conflicto de abril a septiembre de 2007, aunque luego disminuye, tomando su valor más alto como ya se ha indicado anteriormente en agosto de 2009. En enero de 2006 empieza a aparecer un sombreado más intenso, relacionado a tales eventos, que aparecieron con mayor frecuencia. Por tanto, se deduce que en enero de 2006 ocurrió algún evento que desató la escalada hasta alcanzar su punto álgido en agosto de 2009.

9. *Bibliografía:*

• Sanabria, A. M. F., Castañeda, M. P. B., Ramos, R. R. R., and Mateu, J. (2022).

• Identification of patterns for space-time event networks. Applied Network Science, 7(1):1–24. Scott, D. W. (1992).

• Multivariate Density Estimation: Theory, Practice, and Visualization. New York: John Wiley & Sons. Silverman, B. W. (1982).

• Kernel density estimation using the fast Fourier transform. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31(1):93–99. Silverman, B. W. (1986).

• Density Estimation for Statistics and Data Analysis. Routledge. Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019). Spatio-temporal Statistics with R. Chapman and Hall/CRC.

• Mehdi Moradi, Ottmar Cronie, Unai Pérez-Goya & Jorge Mateu (2023): Hierarchical Spatio-Temporal Change-Point Detection, The American Statistician.

• Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. Biometrika, 105(2):455–462.

• Moradi, M. (2018). Spatial and Spatio-Temporal Point Patterns on Linear Networks. PhD Dissertation, University Jaume I.

10. Anexo de código:

• Creación del bandwidth: bw_scott <- lapply(Afghan.pp, bw.scott)

• Media aritmética: bw_ppl_1 <- median(unlist(bw_ppl))

• Función de densidad con la media aritmética para los 72 subconjuntos:

```
int_ppl <- list()
for (i in 1:72) {
 int_ppl[[i]] <- density.ppp(Afghan.pp[[i]],
 sigma=bw_ppl_1,
 leaveoneout = FALSE,
 diggle = TRUE,
 positive = TRUE)}
```

• Creación del ráster:

```
int_scott_raster <- lapply(int_scott, raster)
int_scott_raster <- stack(int_scott_raster)
int_scott_raster <- int_scott_raster*10^9
```

• Mapa obtenido aplicando el ráster:

```
names <- expand.grid(month.name,c(2004:2009))
names(int_scot_raster) <- paste(names[,1],"/",names[,2])
png("Afghanint_scott.png", width = 3800, height = 1660)
spplot(int_scott_raster,
 col.regions=rev(terrain.colors(100)),
 colorkey=list(labels=list(cex=3)),
 scales=list(draw=F),
 par.strip.text=list(cex=3),
 names.attr=paste(names[,1],"/",names[,2]))
dev.off()
```

• Cálculo de la media geométrica:

```
geom_bw_CvL_comun <- do.call(rbind, bw_CvL)
exp(mean(log(geom_bw_CvL_comun)))
geom_bw_CvL_comun = c(exp(mean(log(geom_bw_CvL_comun))))
```

• Cálculo de la función K:

```
dd_scott <- density.ppp(Afghan.pp[[i]], bw.scott, leaveoneout = TRUE) en_scott[[i]] <-
envelope(Afghan.pp[[i]], fun = Kinhom, correction = "border", nsim = 199, nrank = 5, simulate =
expression(rpoispp(dd_scott)), sigma = bw.scott, normpower = 2)
```

• Comparación de mapas: media aritmética y media geométrica:

```
par(mfrow=c(1,2),mar=c(6,2,2,2))

plot(int_scot_raster_geom[[58]], col.regions=rev(terrain.colors(100)),
colorkey=list(labels=list(cex=3)), scales=list(draw=F), par.strip.text=list(cex=3),
names.attr=paste(names[,1],"/",names[,2]) )

plot(int_scot_raster[[58]],col.regions=rev(terrain.colors(100)), colorkey=list(labels=list(cex=3)),
scales=list(draw=F), par.strip.text=list(cex=3), names.attr=paste(names[,1],"/",names[,2]) )
```

• Obtención de las funciones k:

```
png("Función K 2004 c.Scott.png", width = 2000, height = 2000)

par(mfrow = c(4,3),mar=rep(6,4)) for (i in 1:12) { plot(en_scott[[i]], main = "", col="purple", lwd =
3, cex.axis = 2, cex.lab = 2, legendargs = list(cex=2)) }

dev.off()
```