Introduction

Scallops: Scallop data were collected during a 1990 survey cruise off the east coast of North America. Scallop counts were obtained using a dredge. Any scallop smaller than 70 mm was termed a prerecruit. Total catch is the sum of prerecruits and recruits. Measurements included in the data file are:

- National Marine Fisheries Service (NMFS) 4 digit strata designator in which the sample was taken;
- sample number per year ranging from 1 to approximately 450;
- location in terms of latitude and longitude of each sample in the Atlantic Ocean;
- total number of scallops caught at the sample location;
- number of scallops whose shell length is smaller than 70 millimeters;
- number of scallops whose shell length is 70 millimeters or larger.

R*eference: Ecker, M.D., and Heltshe, J.F. 1994. "Geostatistical estimates of Scallop Abundance", In, Case Studies in Biometry, Lange et al., editors. Wiley, New York*

El objetivo de este estudio es analizar los datos "vieiras", identificar su estructura espacial, ya que se trata de datos con referencias geográficas, exponiendo los problemas generales y particulares que surgen durante el análisis. Sugerir diferentes tipos de análisis sobre la aplicación de técnicas y teorías desarrolladas para este propósito. El tema final es predecir los valores dentro de la extensión geográfica del estudio, en los puntos o lugares donde no son observados datos, esto se hará mediante la técnica de predicción espacial descrita en su momento a seguir. Para minimizar la asimetría en los datos (frecuencias o datos de conteo) se utilizó una transformación en el número de vieiras, que se convierte en el logaritmo + 1 (lncatch).

> summary(scp) Sólo variables que se utilizaron en este estudio.

long		lat		lncatch	
Min.:	-73.70	Min.:	38.60	Min. :	0.000
1st Qu.:	-73.15	1st Qu.:	39.44	1st Qu.:	2.197
Median :	-72.73	Median:	39.99	Median:	3.434
Mean:	-72.72	Mean:	39.91	Mean:	3.449
3rd Qu.:	-72.30	3rd Qu.:	40.42	3rd Qu.:	4.736
Max.:	-71.52	Max.:	40.92	Max.:	8.866

> table(scp$strata)

6220	6230	6240	6250	6260	6270	6280	6290	6300	6310	6330	6340	6350
8	16	5	3	12	17	10	5	12	24	10	14	10

Datos “scallops”: log(1 + tcatch) y sus coordenadas:
Latitud (Y Coord) vs Longitud (X Coord)

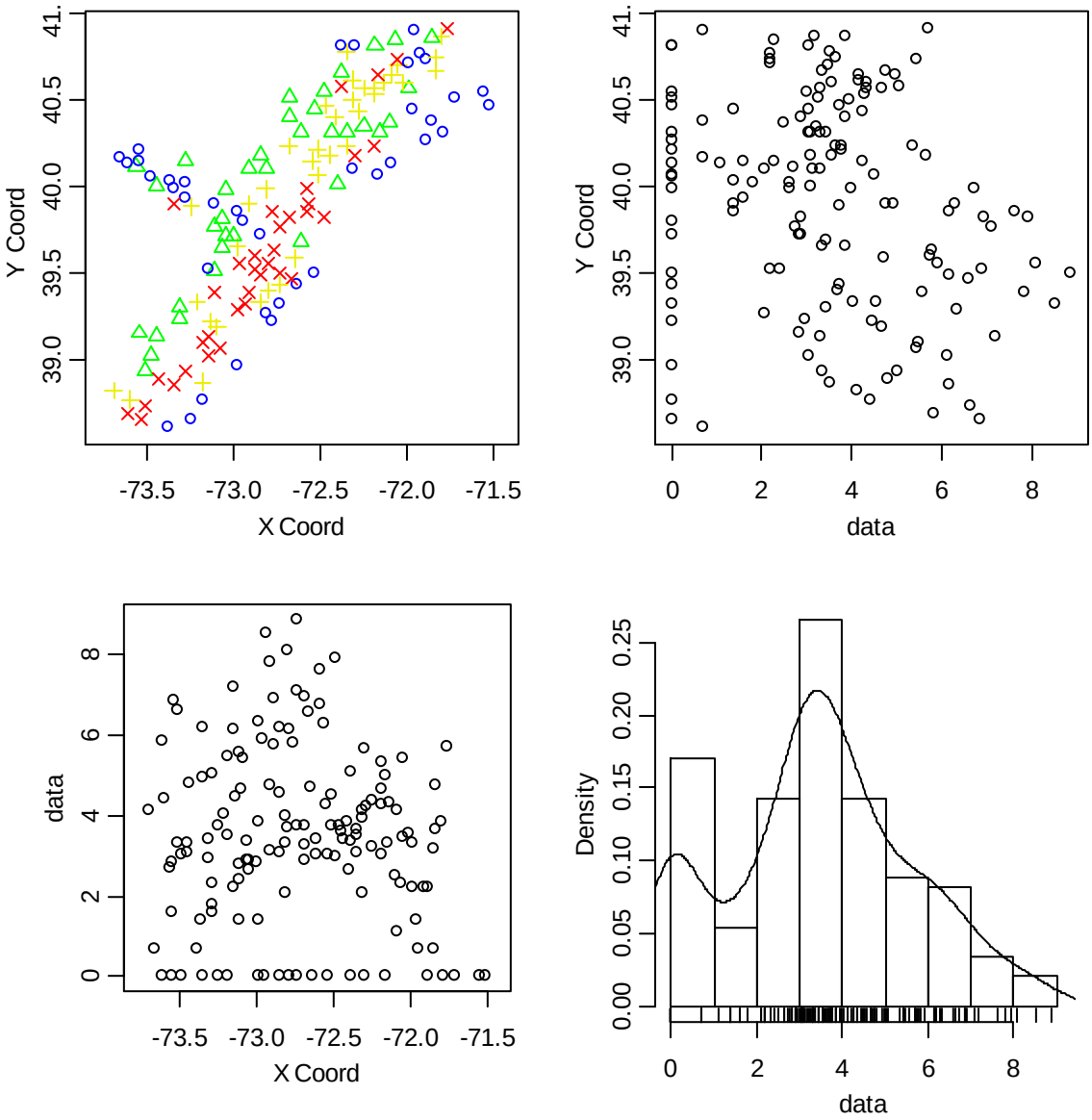

Figura 01: Gráficos producidos por plot.geodata

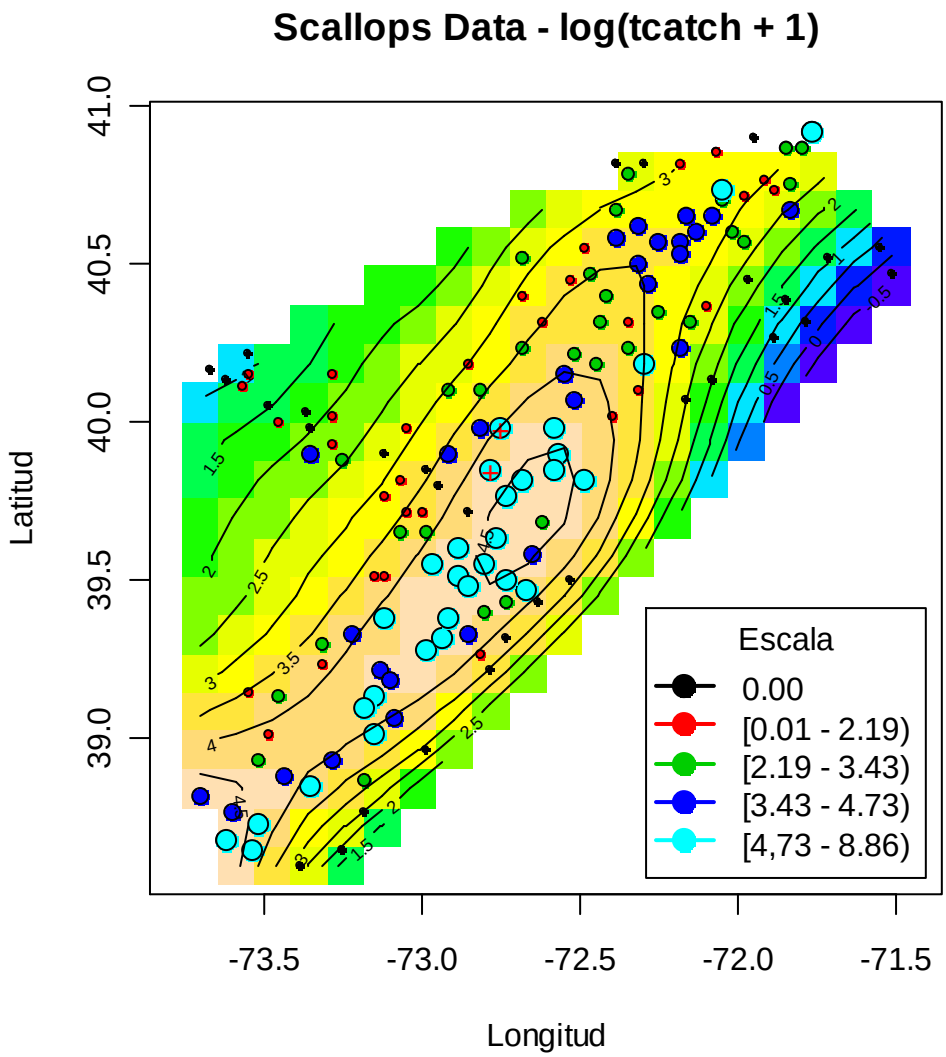

Figura 02: contour, sm regression, image y quantil data scallops

Los datos "Scallops" en el estudio tienen una distribución tal que la mayor cantidad de vieiras es encontrada en la zona centro-sur a lo largo de la este-oeste, de acuerdo con la representación gráfica. También hay un grupo concentrado en azul oscuro en el noreste. La escala utilizada en el gráfico se generó utilizando el método de cuartiles, para los datos transformados "log(tcatch +1)". El color negro indica que la concentración es muy baja (alrededor del 13% de los datos tienen valor cero), y de allí (e incluso el cuartil 25%) son los datos con el color rojo, entre 25% y 50% de los datos es de color verde, del 50% al 75% y el azul oscuro, y entre 75% y el 100% es el color azul.

En este cuadro informativo, las regiones donde se forman las curvas aparentemente homogéneas y cada vez mayor (en forma de elipses) pueden asumir empíricamente que hay presencia de anisotropía geométrica. Ya en los bordes, donde hay unos pocos círculos se asume isotropía. Además tenemos una idea acerca de las tendencias a lo largo de la latitud y longitud, y finalmente los grandes círculos pueden indicar que los valores de datos que son numéricamente más altos que los de las pequeños círculos. Los dos puntos marcados con "+" en el centro son los dos *outiliers* espaciales identificados y eliminados del análisis.

Resumen de los datos

La distribución de los datos es relativamente simétrica, dejando sólo una asimetría en la cantidad de que se aproxima a cero, con una distribución aparentemente normal y sin la presencia de valores atípicos.

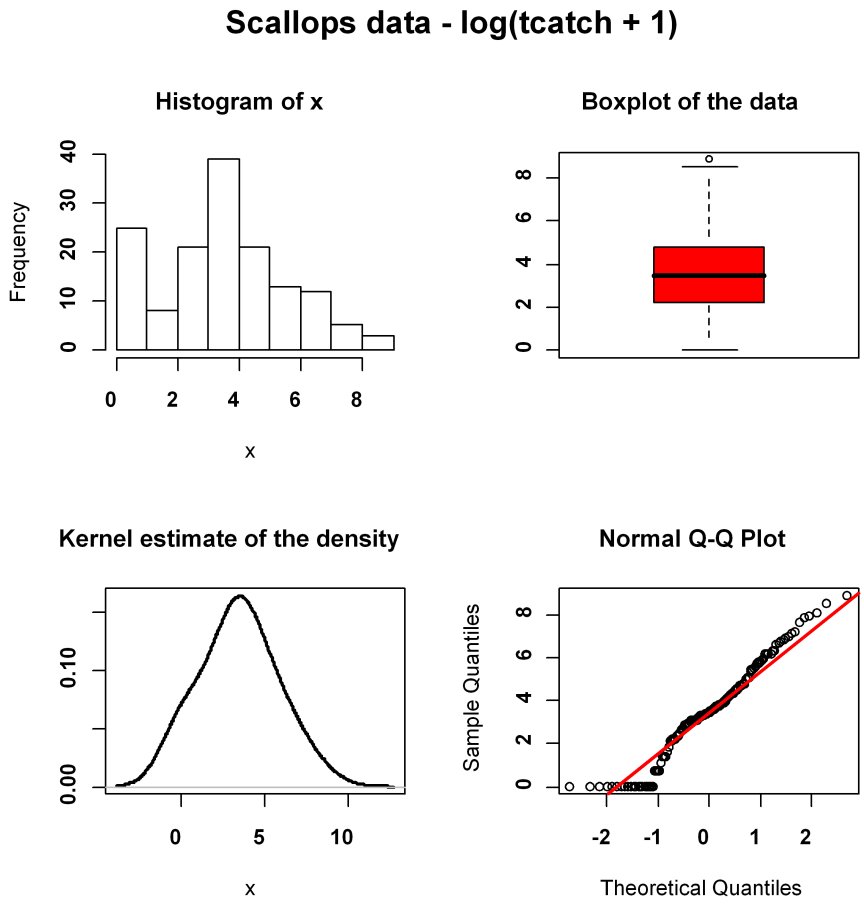

Figura 03: Estadística descriptiva para 'Scallops'

Podemos asumir normalidad, aunque no todos los puntos están bien ayustados a lo largo de la línea teórica da QQ Plot, especialmente a los valores muy pequeños (sobre todo de cero) o muy grandes (7 o más).

Para una evaluación de la variabilidad, se describe una nube de puntos, que son los valores del cuadrado de la diferencia de observaciones en todas las distancias. La nube de puntos es una herramienta de diagnóstico de la posible presencia de valores atípicos o tendencias, también se muestra la variabilidad en relación con el aumento de la distancia entre los datos. Anomalías o falta de uniformidad se observan cuando hay una gran desigualdad en torno a las distancias cortas.

El variograma empírico (Figura 04) es aproximadamente no sesgado como estimador del variograma, mas es estimador poco robusto frente a presencia de valores extremos, que pueden aparecer con cierta frecuencia ya que estas diferencias tienen una distribución con marcada asimetría. La asimetría induce a estimativas sesgadas, lo que genera falsas predicciones. El estimador del variograma empírico clásico es no sesgado cuando el proceso es intrínsecamente estacionario. Sin embargo cuando el proceso es estacionario en segundo orden hay un sesgo que contribuye sustancialmente al error cuadrado medio, cuando n es pequeña.

Variograma empírico que muestra cómo los datos se correlacionan con la distancia.

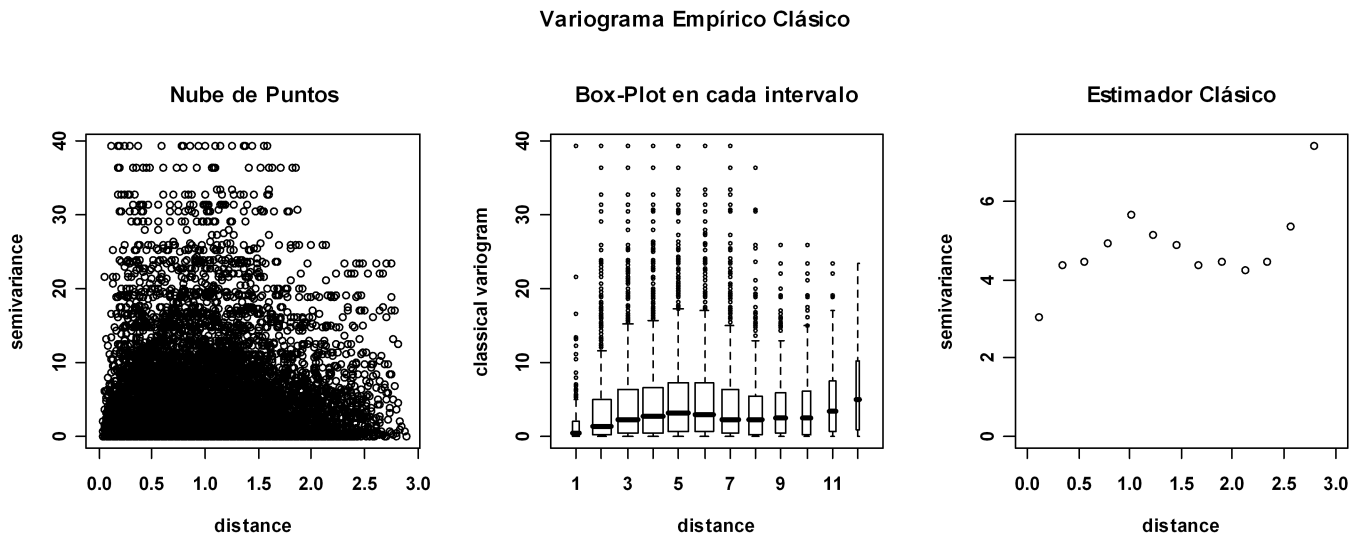

Figura 04: Variograma Empírico Clásico

La variabilidad en distancias cortas parece ser menor que la variabilidad en distancias largas. El estimador clásico para el variograma empírico muestra que a parit de la semivarianza alrededor de 3, la variabilidad es incrementada hasta aproximadamente la distancia 1. Desde entonces disminuye hasta la distancia 2, cuando vuelve a aumentar hasta la distancia máxima de 2.9.

La nube de puntos no parece distribuir al azar, y la mayor concentración de puntos es alrededor de la distancia desde 0 hasta 1,5.

El booxplot de 13 pares (o bins), lo que aparentemente muestra los promedios es la variación a lo largo de la distancia, pero no es claro ya que hay muchos valores extremos de cada par de puntos (bins).

La cantidad de puntos en cada booxplot es:

```
> vieiras.var<-variog(vieiras.geo,estimator.type="classical"); vieiras.var$n
        variog: computing omnidirectional variogram
 [1]  669 1538 1566 1534 1410 1267 1011  724  465  329  218  109   37
```

El variograma empírico robusto no es tan sensible a los valores extremos, los boxplots son más homogéneos, la estimación del variograma es mejor representadas. La vista de alrededor de la nube de puntos es un poco más constante (homocedasticidad) hasta la distancia de 2,5 y la media no son las mismas para todas las distancias. El estimador robusto es muy similar al del estimador clásico.

Es un estimador no sesgado cuando el proceso espacial es gaussiano. Esto podría generar una buena estimación, pero la aplicación directa en geoestadística se descarta en la práctica tanto para estimadores clásicos como para robustos ya que el variograma empírico puede generar varianzas negativas en la predicción espacial por kriging. Esto invalida por completo los resultados porque no son reunidas las condiciones básicas necesarias para un variograma válido.

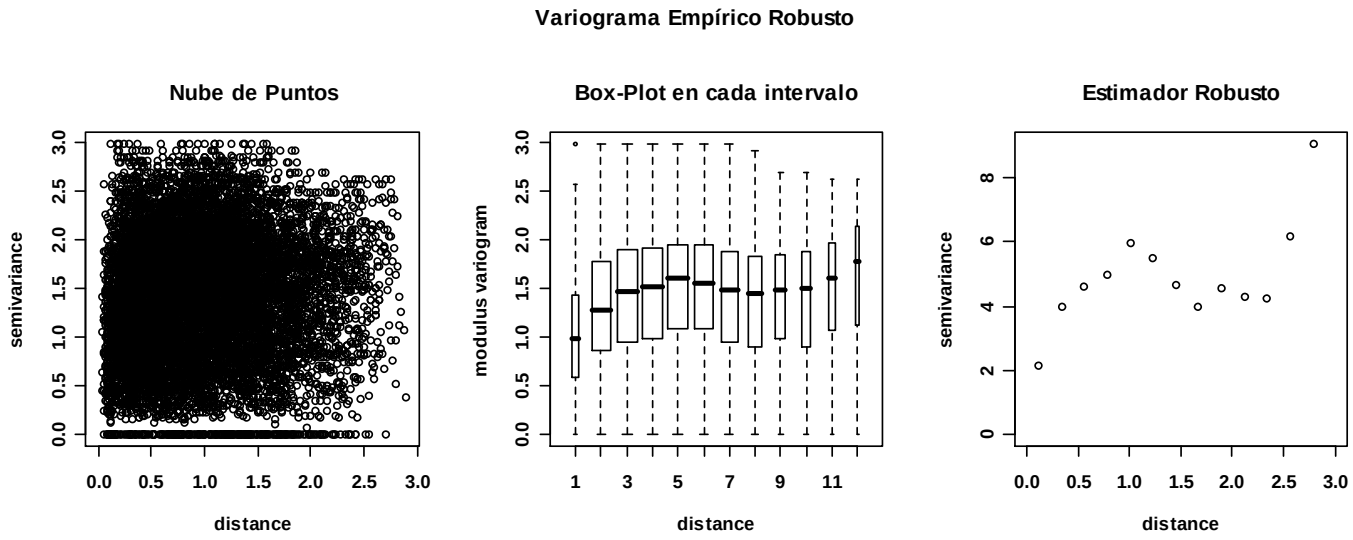

Figura 05: Variograma Empírico Robusto

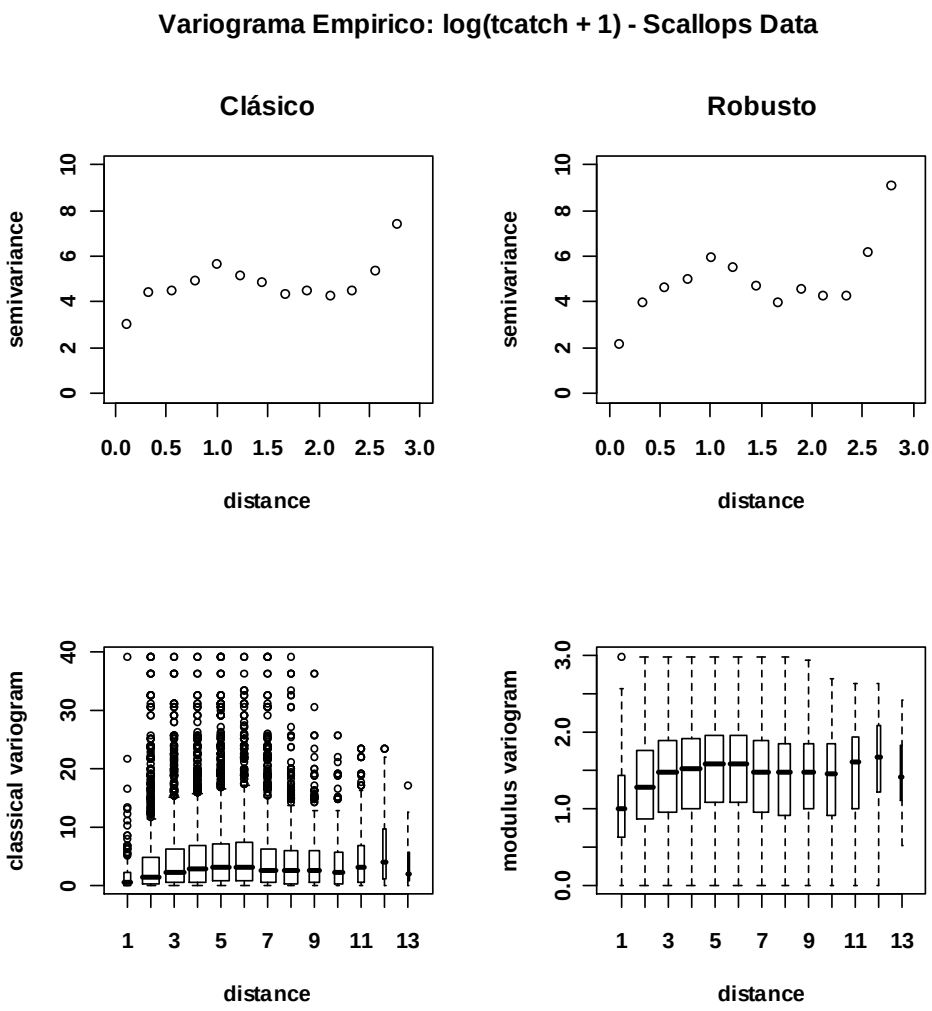

Figura 06: Comparación: Variograma empírico Clásico y Robusto

Se utilizó una distancia euclidiana de hasta 2,9, un total de 13 bins.

La estacionariedad en covarianza se puede representar con las particiones de la área en rectángulos, y se estima un variograma para cada uno de ellos. Para ello se utilizó la variable "*strata*" que está en la base de datos.

.> table(vieiras$strata).

6220	6230	6240	6250	6260	6270	6280	6290	6300	6310	6330	6340	6350
8	16	5	3	12	17	10	5	14	24	10	14	10

Son considerado sólo estratos con más de 6 observaciones. Hay dos posibles valores atípicos en el "*strata*" 6300.

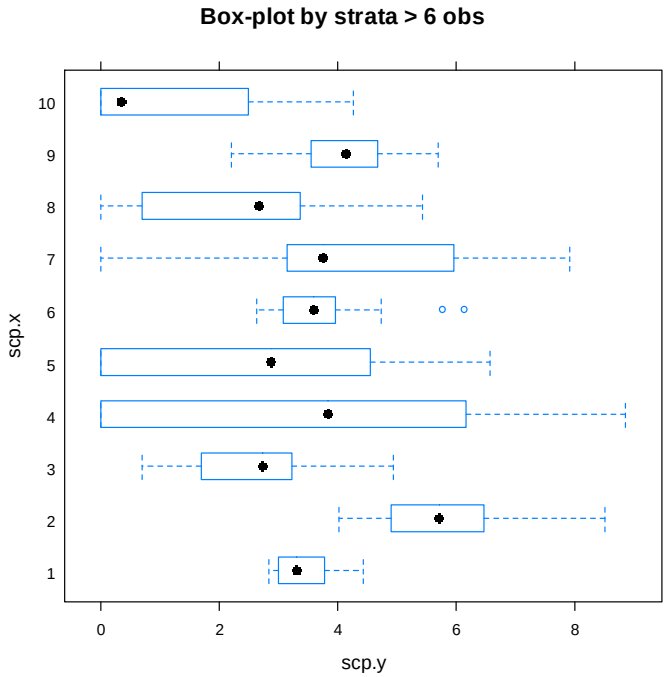

Figura 07 - Estacionaridad en covarianza - 2 outliers espaciais

Sin los *outliers* espaciales (Figura 08).

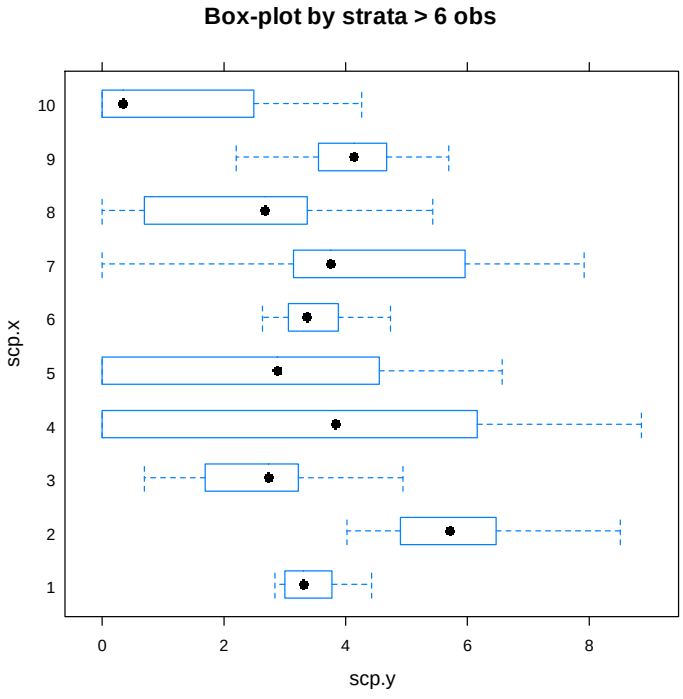

Figura 08 - Estacionaridad en covarianza - outliers removidos

Variograma direccional - para confirmar la isotropía.

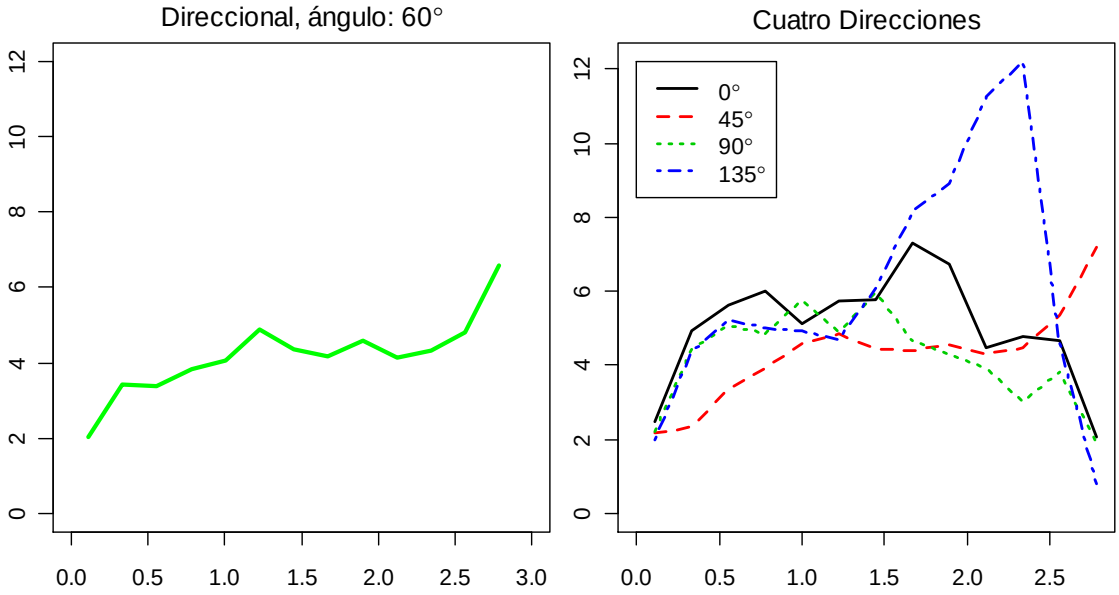

Figura 09: Variograma Direccional a 0º 45º 60º 90º e 135º

Supone que el proceso es isótropo, la dirección a 135 grados es muy diferente de las otras tres, esto puede indicar que hay diferencias a gran escala, ya que varían más en una dirección (con mayor intensidad) que en otras. Pero esta afirmación no es del todo correcta porque en los variogramas direccionales debe ser considerado no sólo las distancias, sino también la magnitud de los datos.

Un método de identificación de anisotropía geométrica es con la visualización de las distancias desde distintas direcciones. Para detectar este tipo específico de anisotropía se utiliza el "Rose Diagram".

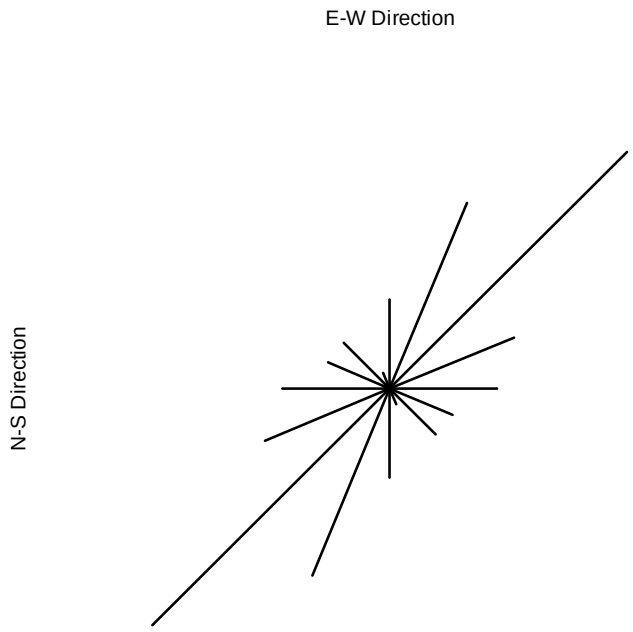

Figura 10: Rose diagram: datos sin rotación

Visualmente se percibe que existe una variabilidad geométrica que se distribuye en 8 direcciones. Las distancias son geométricamente diferentes al cambiar de dirección, alrededor de los bordes forman elipses. Esto indica la presencia de anisotropía geométrica, que puede ser minimizada.

La rotación de los ejes nos permite visualizar mejor el efecto geométrico.

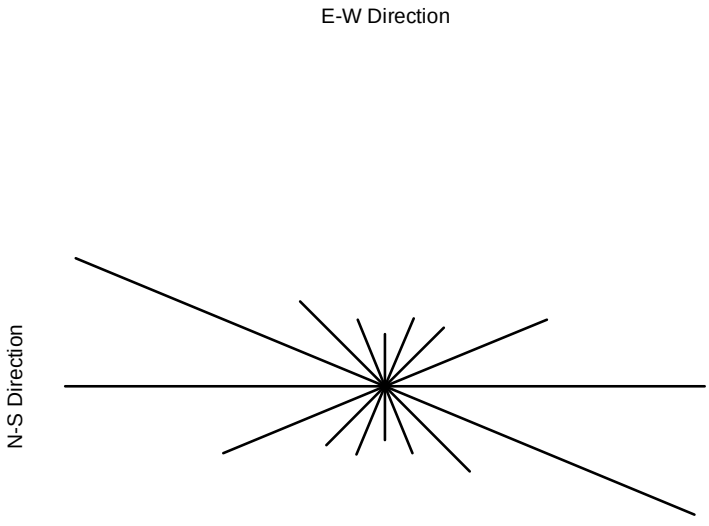

Figura 11: Rose diagram: datos con rotación a 52º

Ajuste del modelo paramétrico de variograma (con los datos originales).

El modelo elegido para el Variograma Teórico fue el modelo Exponencial con la estimación de parámetros por el método de máxima verosimilitud restringida, REML, bajo el supuesto de que el proceso es gaussiano, este método estima los parámetros simultáneamente con algoritmos numéricos internos para la convergencia mutua. Por otra parte, los métodos REML son menos sesgados que MLE y otros métodos.

Varias pruebas se realizaron en un intento por ajustar algunos otros modelos (“a ojo”, por mínimos cuadrados, etc.) pero el resultado visto a través de envolventes de variogramas empíricos no fueron satisfactorios en comparación con REML, bajo la hipótesis de independencia espacial.

Modelo REML corrige el efecto de la anisotropía y corrige la tendencia con estimaciones de los parámetros mutuamente (“Parameters of the mean component” (trend): beta 2.2261). Se interpreta, por el modelo REML, que el modelo ajustado es isotrópico, ya que lo resultado es:

```
> reml<-likfit(scp.geo, ini=pars, fix.psi=F, fix.psiR=F, fix.nugget=F,
                    nugget=nug, lik.method="REML", cov.model="exp")
> summary(reml)
…
anisotropy parameters:
    (fixed) anisotropy angle = 0  ( 0 degrees )
    (fixed) anisotropy ratio = 1
```

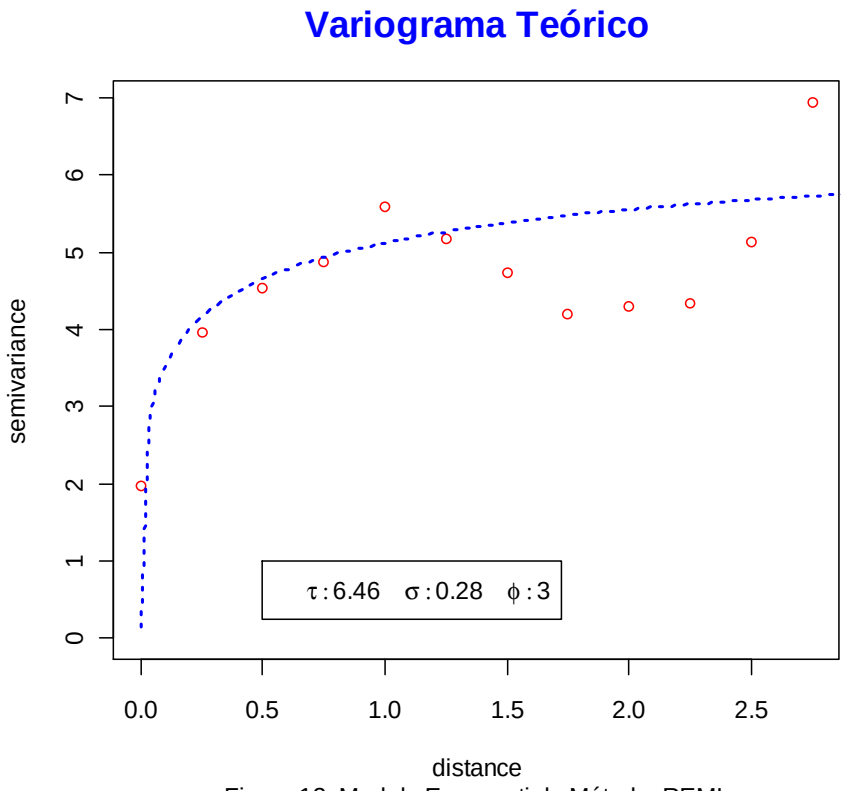

Figura 12: Modelo Exponential - Método: REML

El variograma utilizado en la estimación "*Kriging*" para los datos originales (Figura 12). La distancia máxima utilizada es alrededor de 2.9, en la que los siguientes parámetros fueron estimados por el modelo de variograma teórico: "*Exponential*" y el estimación por el método REML - con la función *likfit* (con el supuesto de normalidad del proceso). Teniendo en cuenta que el efecto pepita se estima muy por debajo de la semivarianza 2 correspondiente al primer bin.

Uso de modelo aditivo generalizado - función gam(.)

Se ajustó modelo a las nuevas coordenadas, que mostraron mejores resultados que el modelo estándar (sin rotación de ejes). Con la rotación, las variables "longitud" y "latitud" son muy significativas para el modelo, las estimaciones fueron mejores, y el valor de la "*deviance explained*" aumentó de 17% (modelo estándar) para 58% (modelo con rotación de ejes).

Modelo con Rotación > summary(gam.scp.rot)

```
> summary(gam.scp.rot)

Family: gaussian
Link function: identity

Formula:
lncatch ~ s(long) + s(lat)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4485     0.1204   28.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
          edf Ref.df     F  p-value
s(long) 2.208  2.208 7.178 0.000727 ***
s(lat)  7.572  7.572 22.043  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.55   Deviance explained =   58%
GCV score = 2.2837  Scale est. = 2.1151    n = 146
```

Modelo Sin Rotación > summary(gam.scp)

```
Family: gaussian
Link function: identity

Formula:
lncatch ~ s(long) + s(lat)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4485     0.1666    20.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(long) 2.664  2.664 3.425  0.0233 *
s(lat)  2.700  2.700 2.624  0.0590 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.138   Deviance explained = 16.9%
GCV score = 4.2369  Scale est. = 4.0522    n = 146
```

La eliminación de la tendencia lineal (NE-SW), con la costa hacia el interior se hizo de la aplicación de los residuos generados por el modelo aditivo generalizado, con la estimación por la función gam() en "R", y rotación de 52º en los ejes. Esto elimina el efecto de la tendencia en los datos.

Plan de tendencia antes y después de la rotación.

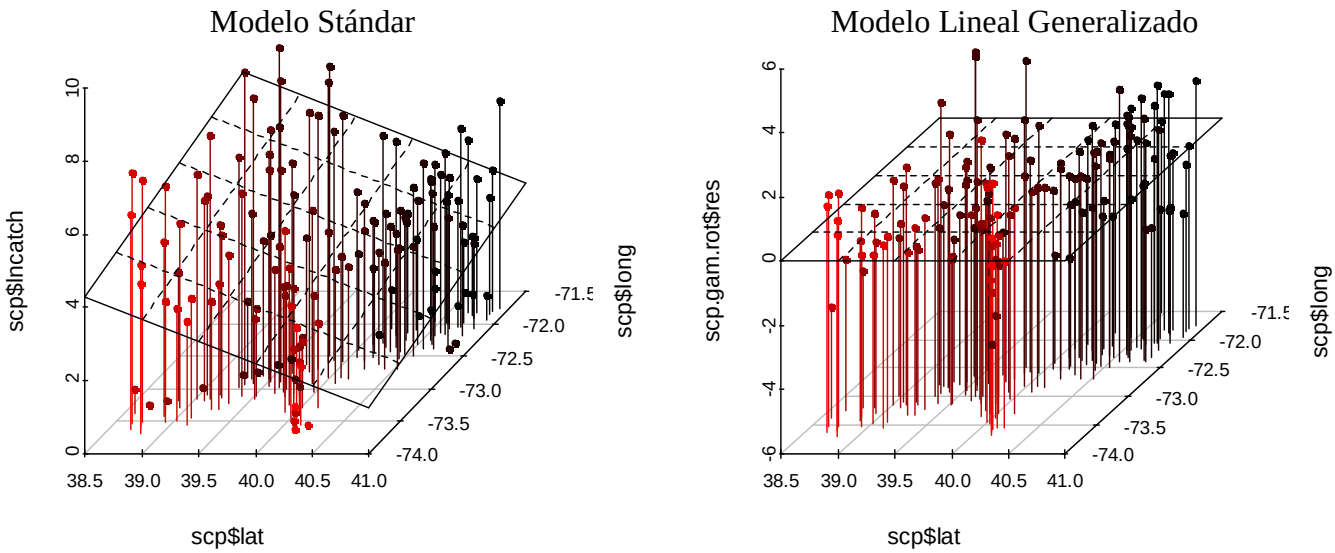

Figura 13: Plan de tendencia en los datos originales, y el modelo GAM con rotación en los ejes.

El variograma generado en las cuatro direcciones (Figura 14) tiene más homogeneidad, lo que causa estacionariedad en la covarianza, que puede ser comprobada por variogramas en diferentes direcciones.

De acuerdo con los variogramas direccionales a continuación hay menos variabilidad entre las diferentes direcciones (la semivarianza es entre 1 e 3). Todavía un poco de la variabilidad indica que el valor intrínseco a la magnitud de los propios datos son los que ponen más peso a la variabilidad. El efecto de la media, que aparentemente era inestable, fue eliminado.

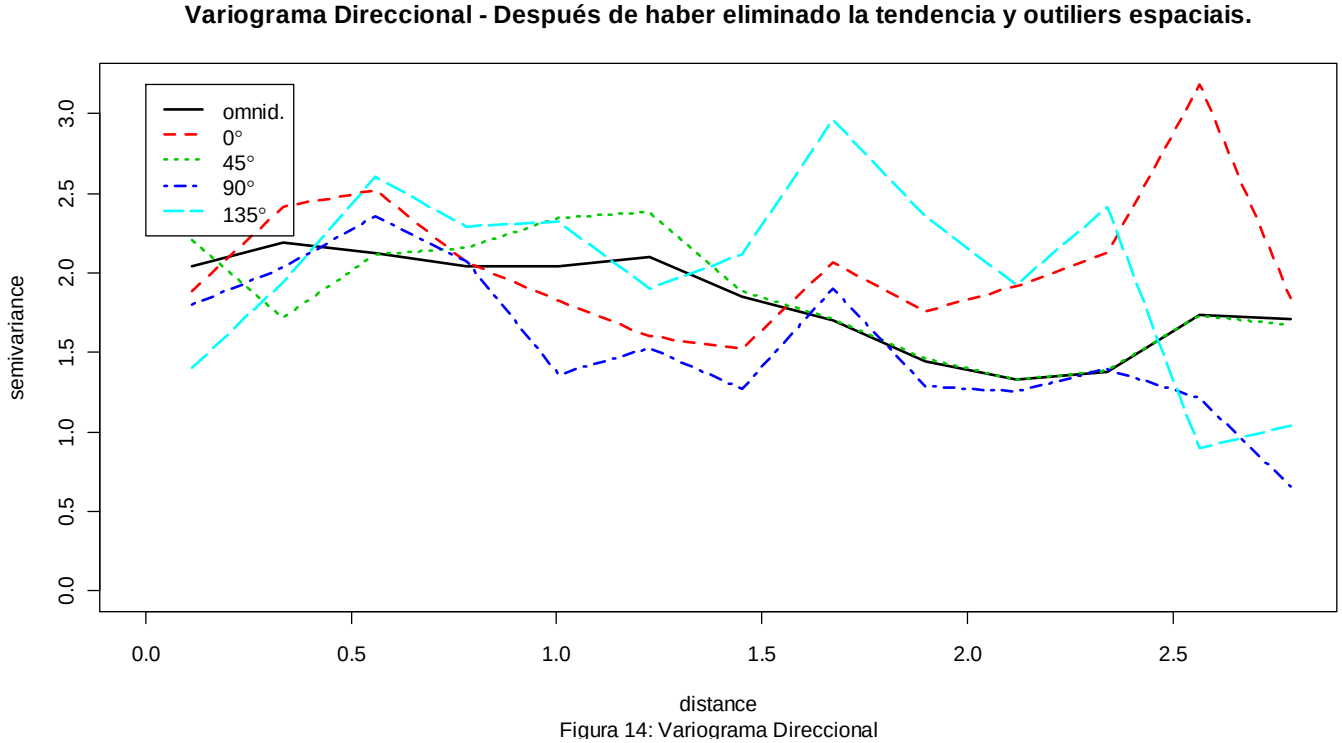

Figura 14: Variograma Direccional

Los gráficos siguientes (Figuras 15) son de los residuos del modelo ajustado con el nuevo eje y sin *outliers* espaciais. Hay distribución simétrica y normal con media cero.

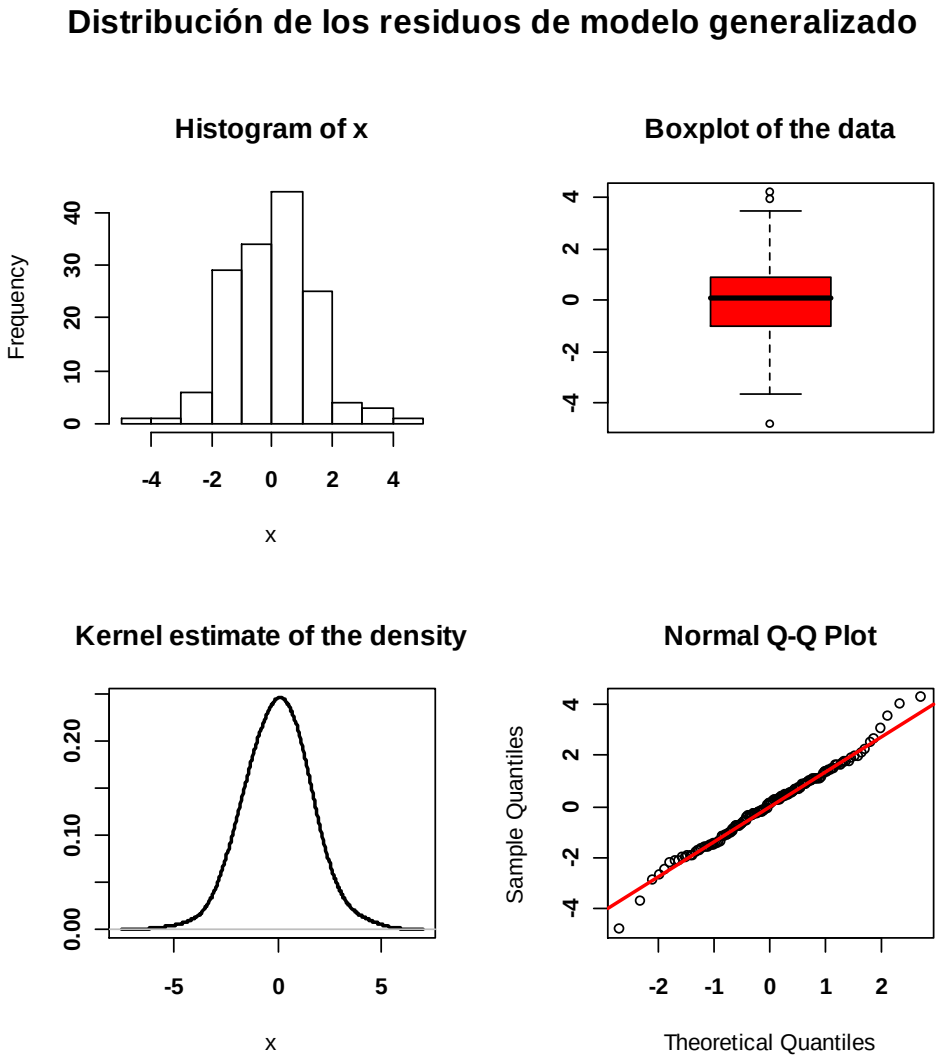

Figura 15: gráfico de residuos para el modelo: gam(lncatch ~ s(long) + s(lat))

Los residuos generados fueron utilizados para volver a calcular el variograma (Figuras 16 y 17) según el modelo:

$$\mu = 3{,}4485 + 2{,}208 \times s(\text{longitud}) + 7{,}572 \times s(\text{latitud})$$

desde el ajuste del modelo generalizado por medio de la rotación de los ejes.

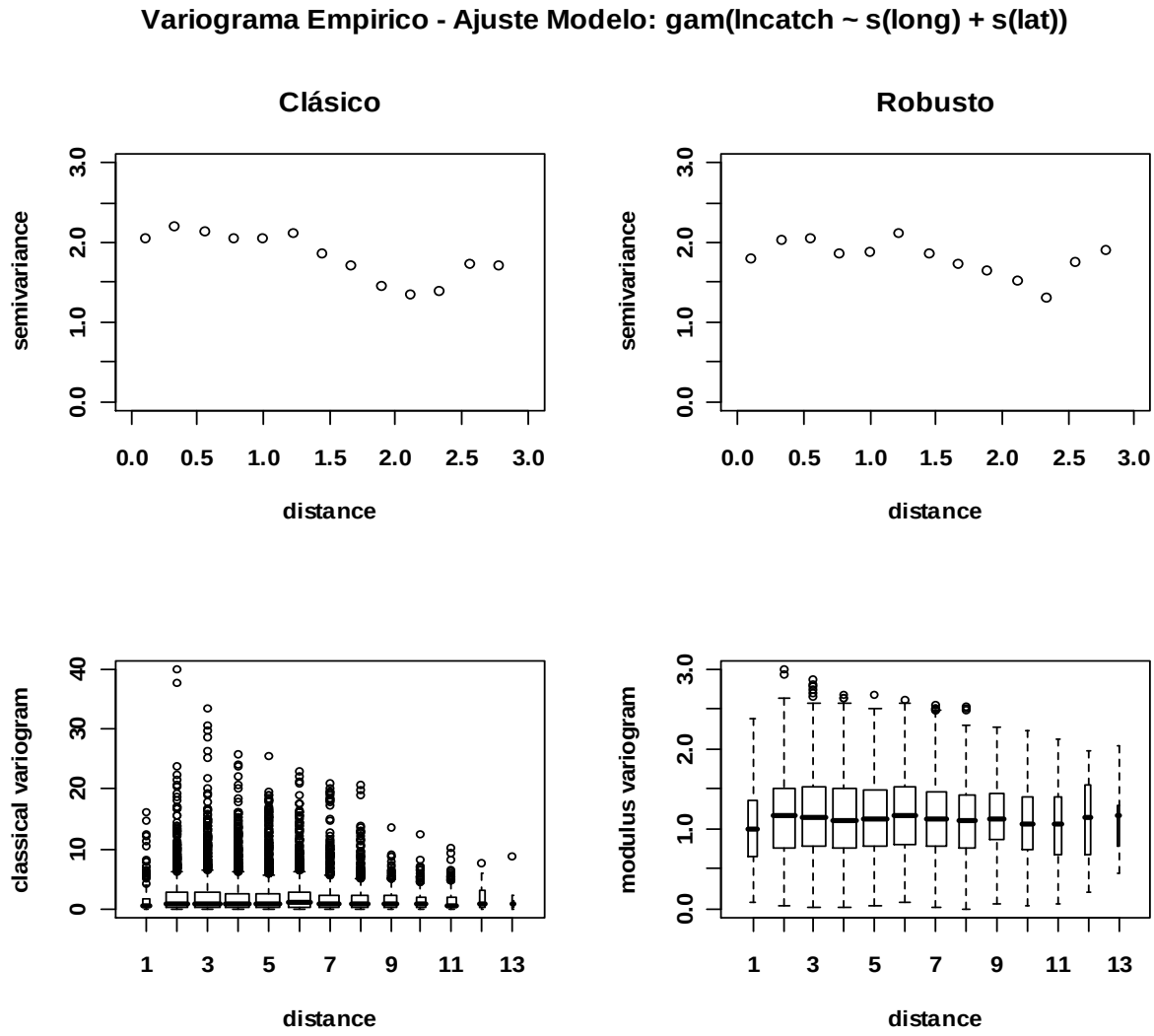

Figura 16: Comparación: Variograma Empírico Clásico y Robusto

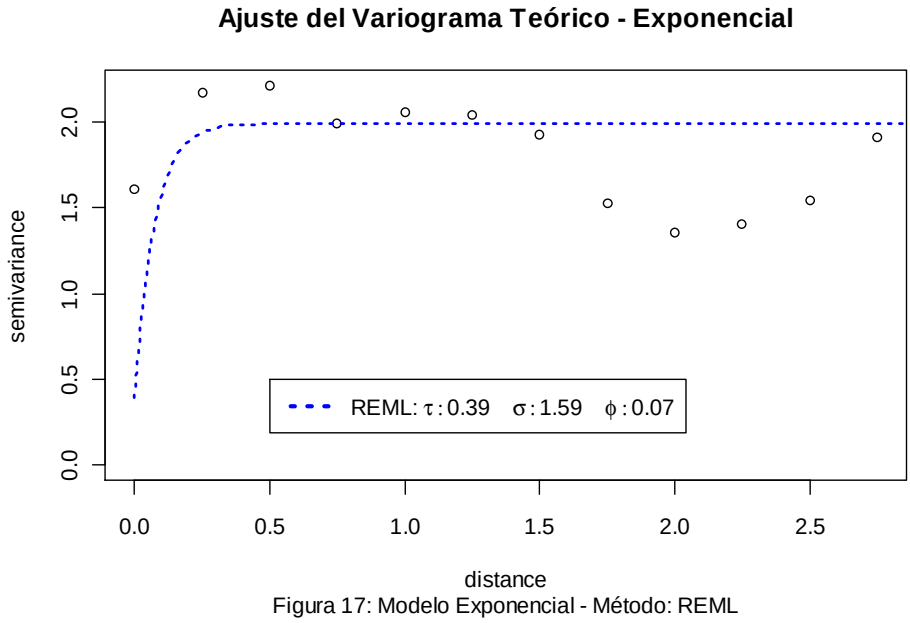

Figura 17: Modelo Exponencial - Método: REML

Para el modelo "Exponencial", los parámetros de variograma se estiman por el método REML, teniendo en cuenta la distancia máxima de 3 para el variograma. La semivarianza es de alrededor de 2 y la media del proceso es cero. El primer "bin" del modelo anterior (Figura 12) se inicia en semivarianza 2 (nugget).

Se estableció una malla regular con 51x51 puntos para las predicciones realizadas con Kriging.

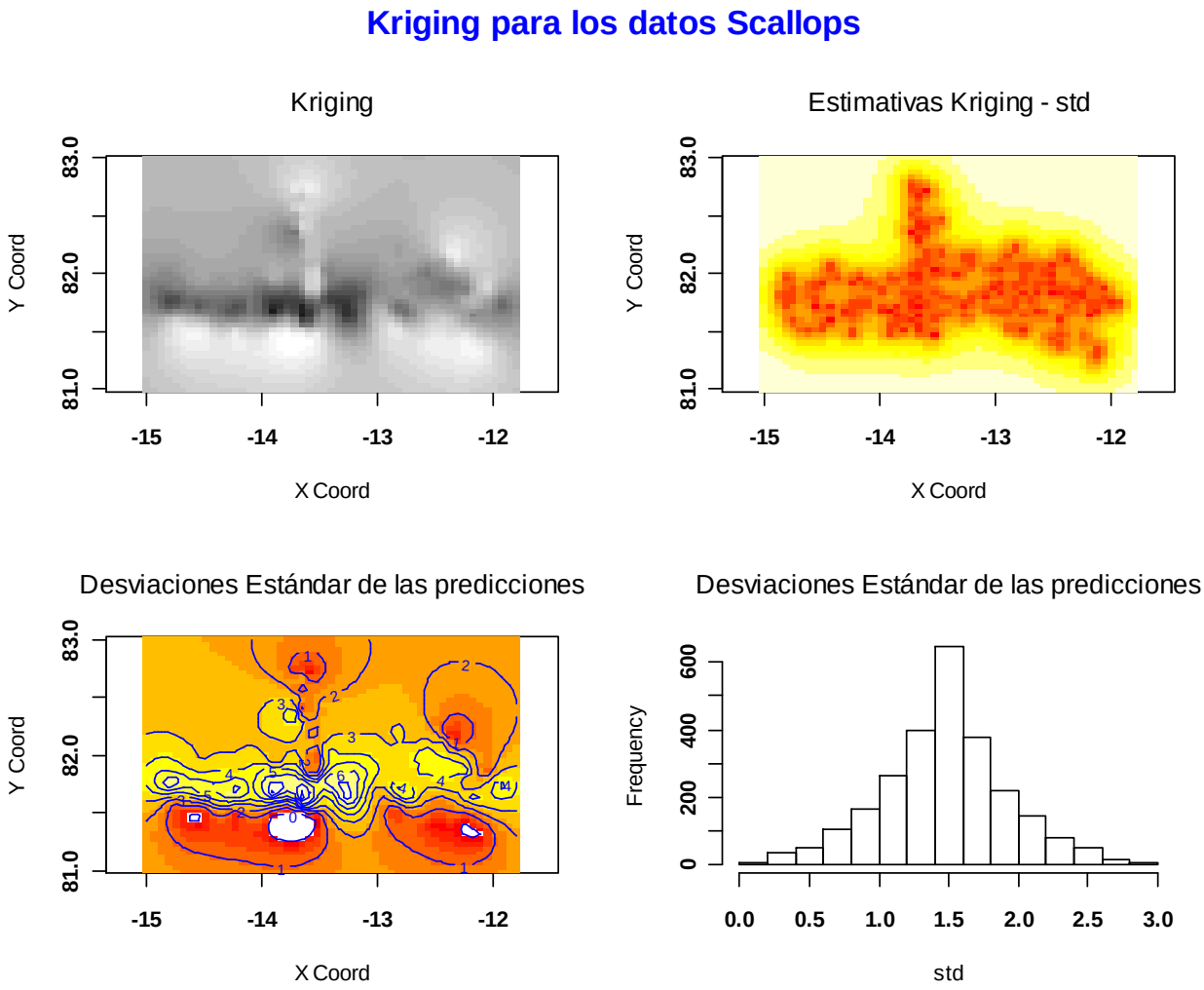

Figura 18: Kriging para los datos originales

En el primer gráfico, las “manchas” oscuras muestran los sitios donde hay mayor concentración de *scallops*. Son los más altos valores de los datos, teniendo en cuenta la malla cuadrada de relieve rectangular.

El segundo gráfico muestra la desviación estándar de las estimaciones realizadas por el método de *Kriging*. Tenga en cuenta que es bastante uniforme en toda la región donde se observaron los datos. Es más claro en los bordes, lo que significa una desviación mayor en las regiones distantes del sitio de las observaciones reales.

El tercero gráfico muestra la desviación estándar de las predicciones - "suavizado" por el modelo generalizado donde los colores en rojo son aquellos con la menor desviación, donde es más homogénea la cantidad de *scallops*. En la región seleccionada que muestra los blancos la desviación es cero, representan un sitio donde la predicción es más homogénea, probablemente casi no hay datos por allí.

El último gráfico indica la distribución de las desviaciones de las predicciones.

La siguiente Figura 19 muestra un polígono ajustado de acuerdo a los datos observados. (Ver las coordenadas en el ANEXO I)

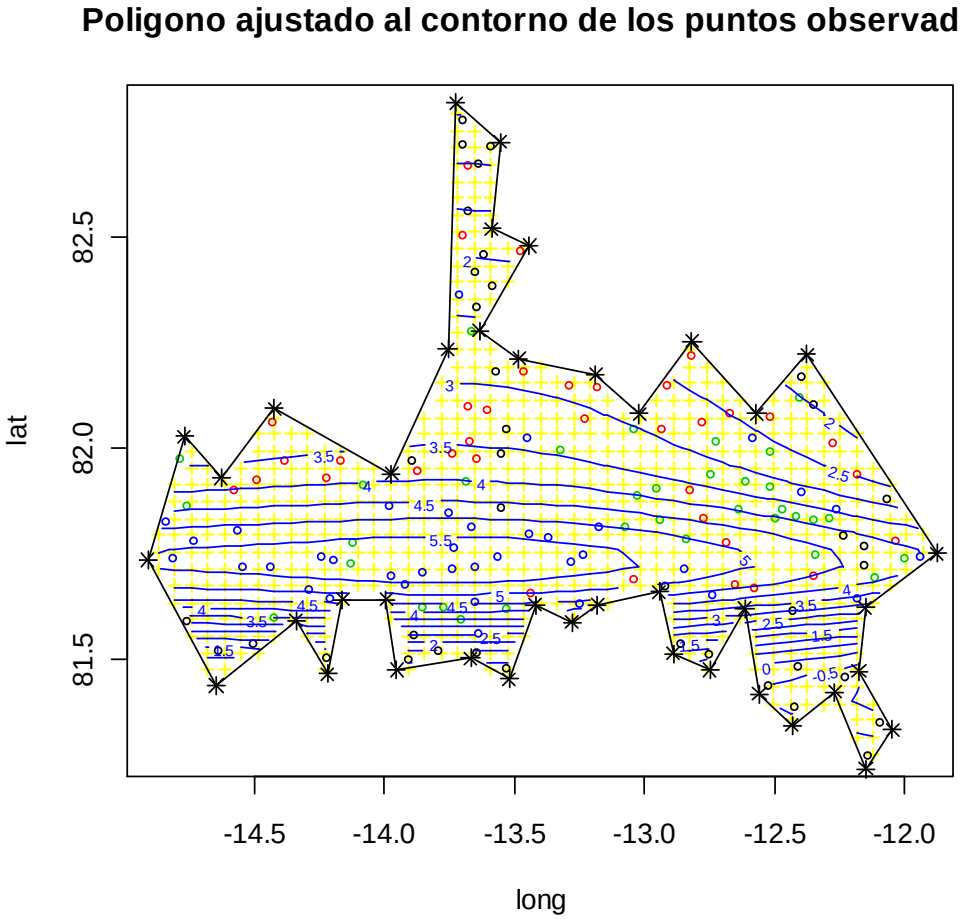

Figura 19: Región de predicción de acuerdo a los datos observados

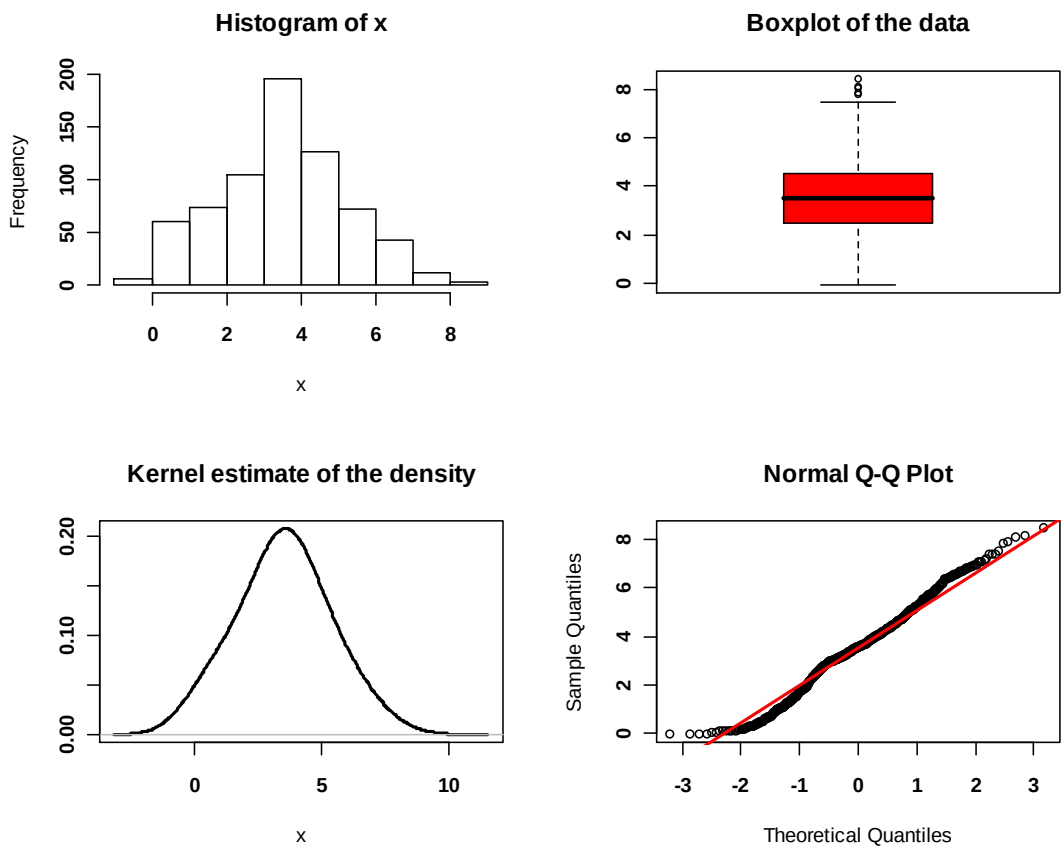

Figura 20: Predicciones con el polígono ajustado a los datos

Mediante la adaptación de la región de predicción (Figura: 19), de acuerdo con las observaciones, se puede verificar que las estimaciones son mejores (Figura: 20), algo que puede ser "probada" visualmente en el gráfico de la densidad relativa e por el QQ-plot, donde los puntos predictos se ajustan bien al cuantil teórico.

Los "envelopes (o sobres) simulados" (Figura: 21) son modelos que se construyen a partir de simulaciones. Basado en un determinado conjunto de parámetros del modelo, que en este caso es el modelo REML para *Kriging* Ordinario y OLS para *Kriging* Universal. Las simulaciones se calculan mediante la generación de valores en cada sitio de acuerdo con un modelo de Gauss, con el modelo de variograma elegido y con los parámetros estimados. Para cada una de las simulaciones, calcula el variograma empírico, utilizando los mismos “bins” para el variograma empírico de los datos. Los “envelopes” fueron calculados teniendo para cada “bin” el máximo y mínimo del variograma empírico en los datos simulados.

El "envelope" a continuación (Figura 21) son para los dos casos de *Kriging*, que apoyan la hipótesis de independencia, puesto que ningún “bin” está fuera de límite superior y inferior del "envelope", según el modelo utilizado, mientras que en el segundo “*Kriging*” el efecto de la media fue eliminado. Se puede concluir que hay independencia espacial en ambos casos.

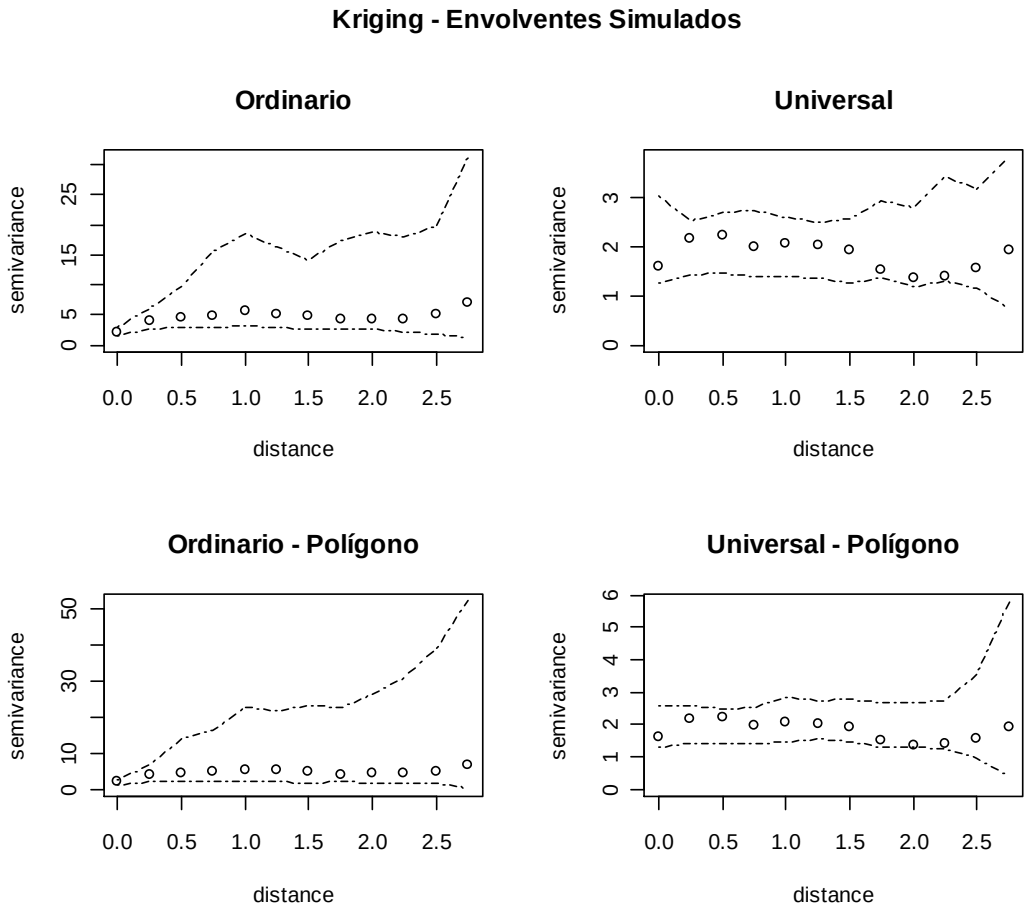

Figura 21: Envolventes Simulados para Kriging Ordinario y Universal

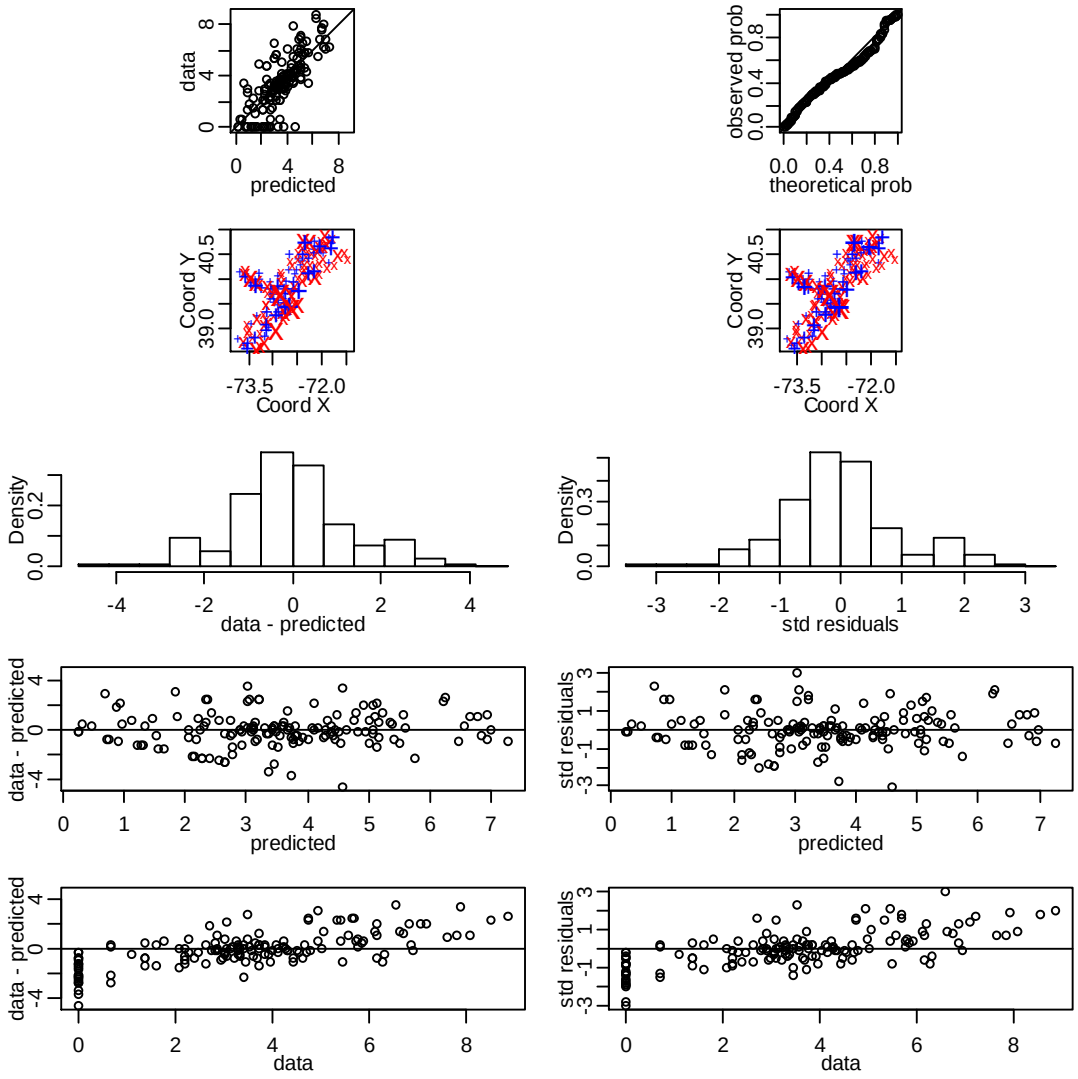

Figura 22: Cross Validation - Kriging Ordinario - Metodo REML

En el primer gráfico, los puntos son distribuidos casi en su totalidad en la línea, los valores que están por debajo son los que antes de la transformación de log (tcatch +1) fue igual a cero y siguen siéndolo.

En el segundo gráfico (qqplot) hay alguna desviación, que no es mucho, lo que indica cerca de lo normal. Hay muchas observaciones que se distribuyen entre las desviaciones -1 y 1, que es bueno y prácticamente no hay nada fuera de los rangos de desviación -3 y 3.

Ambos histogramas, la distribución de los valores preditos por la desviaciones estándares de los residuos son básicamente simétricos. Los dos gráficos con el eje de valores preditos tienen una distribución casi simétrica sobre la línea, lo que indica poca variación en pequeña escala. Mayor observacion, mayor es el residuo, esto es una señal a la magnitud natural de los datos.

En general, una gran parte de los valores dan una buena estimación.

Los últimos gráficos muestran una pequeña tendencia, ya que algunos puntos están por debajo de la línea.

El análisis de la validación cruzada para la predicción por *Kriging* Universal es básicamente idéntico a la *Kriging* Ordinario.

Decisión sobre la validación cruzada. (VC1, VC2 y VC3)

```
xv.ko  <- xvalid(scp.geo,model=reml)              # son los datos originales
xv.ku  <- xvalid(scp.geo.gam.res,model=reml)  # los datos son los residuos del ajuste gam(.)

> reml
likfit: estimated model parameters:

   beta    tausq  sigmasq      phi     psiA     psiR
"2.2261" "0.1470" "6.4587" "0.2754" "0.0000" "1.0000"
Practical Range with cor=0.05 for asymptotic range: 0.8251119

########################## KRIGING ORDINARIO
 > n<-dim(scp)[1]

> VC1 <- (1/n)*sum(-xv.ko$error/sqrt(xv.ko$krige.var));VC1
[1] 0.01671149

> VC2 <- sqrt((1/n)*sum((xv.ko$error/sqrt(xv.ko$krige.var))^2));VC2
[1] 0.9767928

> VC3 <- sqrt((1/n)*sum(xv.ko$error^2));VC3
[1] 1.396681

> #### VC1=0.0167, VC2=0.9768, VC3=1.3967


########################## KRIGING UNIVERSAL
> n<-dim(scp)[1]

> VC1 <- (1/n)*sum(-xv.ku$error/sqrt(xv.ku$krige.var));VC1
[1] 0.002412397

> VC2 <- sqrt((1/n)*sum((xv.ku$error/sqrt(xv.ku$krige.var))^2));VC2
[1] 0.9921726

> VC3 <- sqrt((1/n)*sum(xv.ku$error^2));VC3
[1] 1.291263

> #### VC1=0.0024, VC2=0.9922, VC3=1.2912
```

Las medidas VC1, VC2 y VC3 indican que el Kriging Universal es mejor. VC3 y VC1 son más pequeños y VC2 está más cerca de 1.

Conclusiones Finales

El uso del software R se ha convertido en indispensable para obtener estos resultados, se hicieron numerosos intentos para adaptar las técnicas y teorías sobre el problema, así que adaptaron a los datos más allá de la familiarización de los comandos del paquete R.

Los datos *scallops* han sido ampliamente explorados en los manuales de estudio de estadística espacial, de alguna manera siempre contribuyen en gran medida a la lectura de diversas fuentes con diferentes aplicaciones que son a menudo con imposiciones demasiado largas, o difícil de entender. Creo que es un trabajo muy útil en la diversidad y complejidad que nos presenta la Geoestadística, sobre todo porque es una técnica estadística reciente pero en constante avance científico y académico.

ANEXO I – Coordenadas Xpoly - Coordenadas que se utilizaron para ajustar el polígono.

> Xpoly

	x	y
1	-13.72478	82.81983
2	-13.55524	82.72548
3	-13.58806	82.52105
4	-13.44313	82.48016
5	-13.63454	82.27887
6	-13.48415	82.21283
7	-13.18883	82.17509
8	-13.01930	82.08073
9	-12.81696	82.25371
10	-12.56813	82.08073
11	-12.37398	82.22226
12	-11.87359	81.75050
13	-12.14976	81.62155
14	-12.17437	81.47059
15	-12.04585	81.33220
16	-12.14703	81.23785
17	-12.26734	81.42027
18	-12.43141	81.34164
19	-12.55719	81.41712
20	-12.61461	81.61841
21	-12.74860	81.47373
22	-12.88805	81.51147
23	-12.94274	81.65929
24	-13.18063	81.62784
25	-13.27907	81.58696
26	-13.42126	81.62784
27	-13.51970	81.45172
28	-13.67009	81.50204
29	-13.95447	81.47373
30	-13.99548	81.64042
31	-14.16228	81.64042
32	-14.21970	81.46430
33	-14.34002	81.59010
34	-14.64900	81.43599
35	-14.91151	81.73477
36	-14.77205	82.02727
37	-14.62986	81.92977
38	-14.42478	82.09331
39	-13.97361	81.93921
40	-13.75759	82.23484

>