Desplazamiento de la onda

Esto surgiócon Carolina mientras discutíamos el tema de la estacionariedad de las ondas. Van Loon 1972 tiene este gráfico donde grafica el cambio de fase diario de la onda en función de la amplitud. Lo que ve es que cuando cambia poco (o nada) la amplitud es más alta. Esto sería una definición directa de “onda estacionaria” con la limitación de que el cambio de fase en 24hs de una fase de Fourier no es necesariamente lo mismo que la velocidad de fase de la onda en sí.

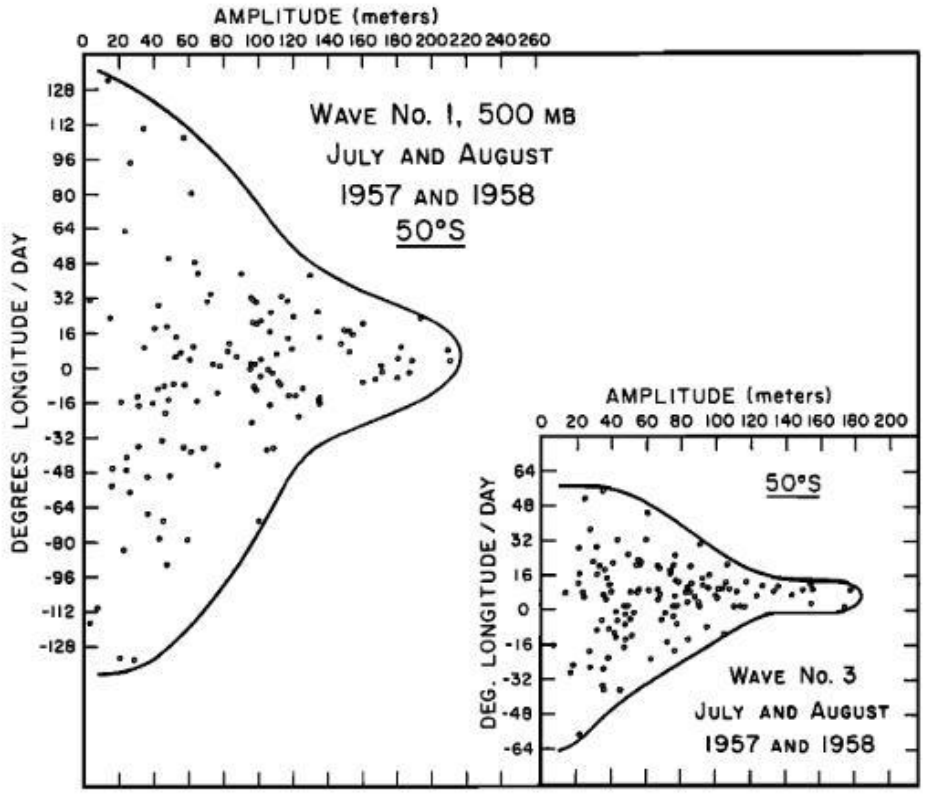
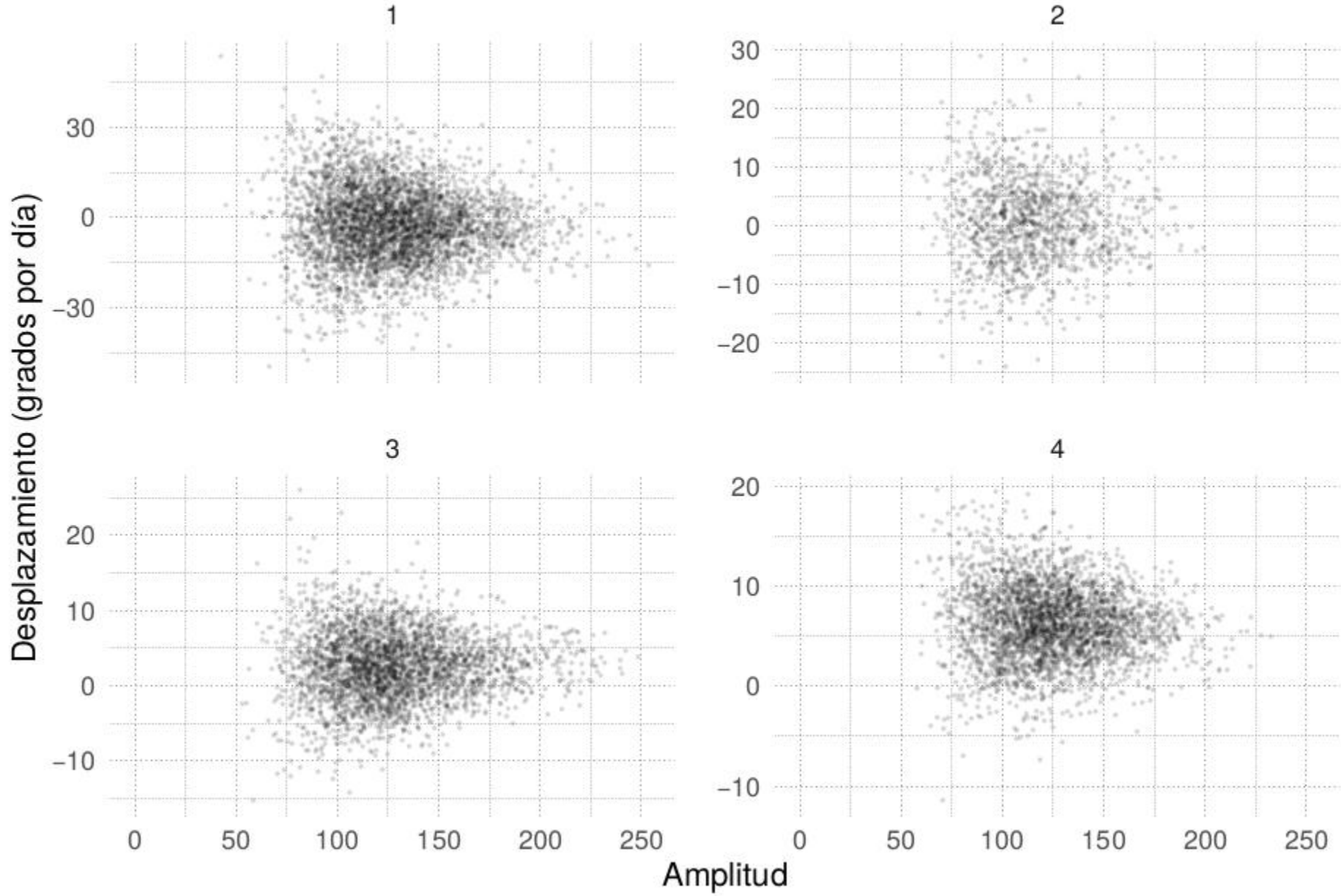
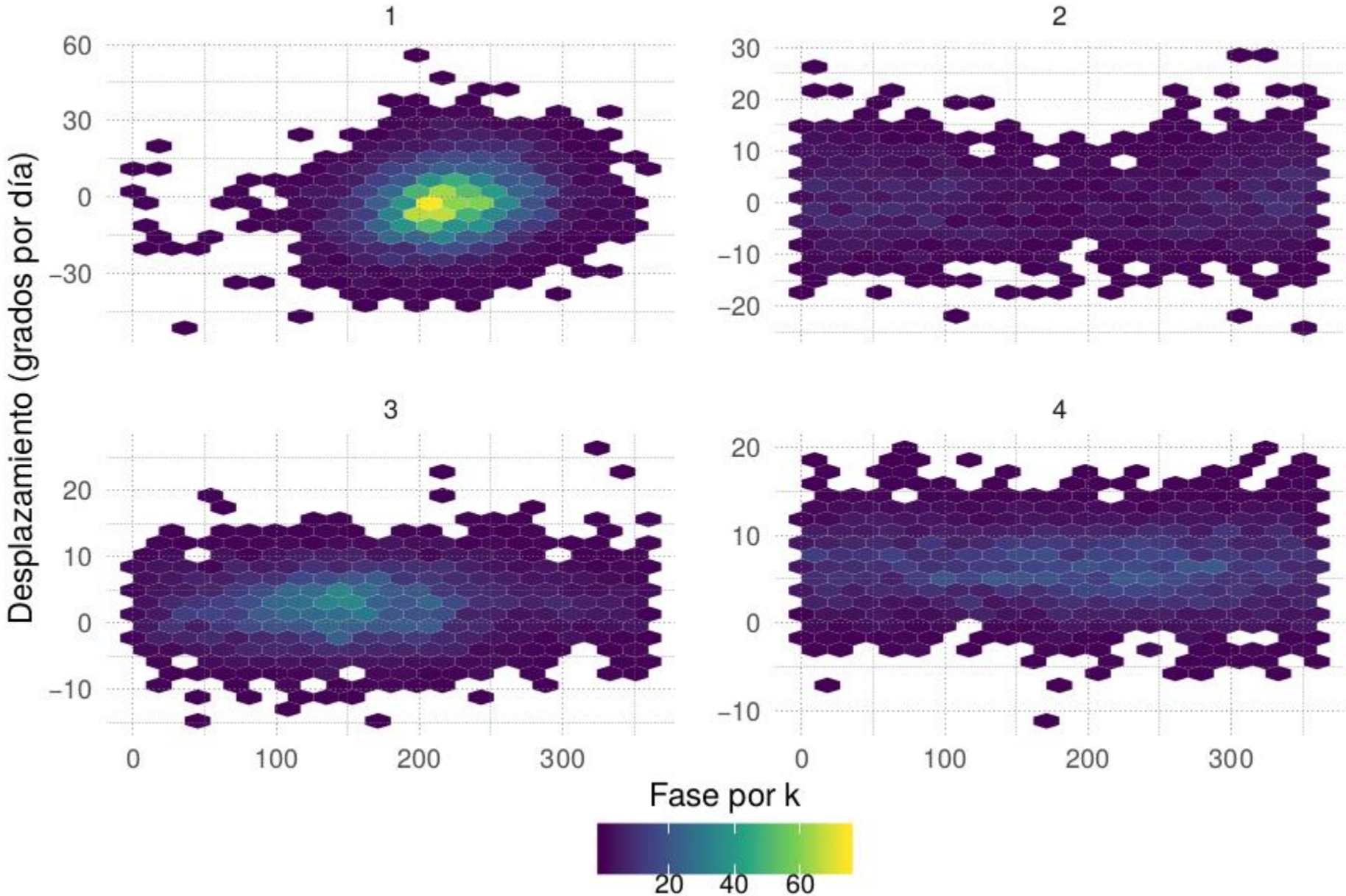


Fig. 13. Daily amplitudes (meters) of waves 1 and 3 at 500 mb in winter plotted as a function of 24-hour phase changes.

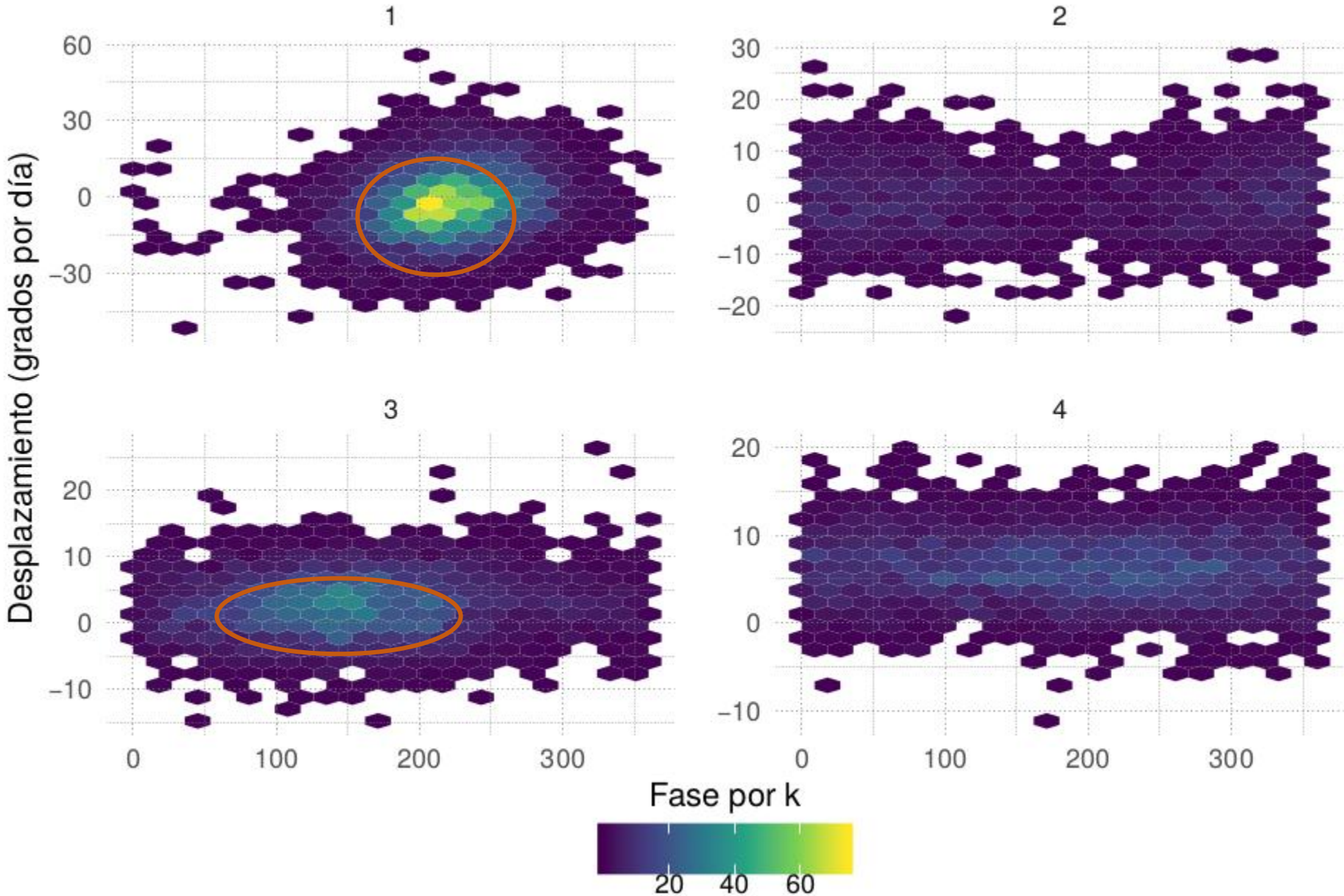
van Loon, H., & Jenne, R. L. (1972). The zonal harmonic standing waves in the southern hemisphere. Journal of Geophysical Research, 77(6), 992–1003.



Amplitud vs. desplazamiento para k = 1-4 para casos donde el r^2 de la onda es mayor a 0,25



Suma de la amplitud normalizada (en colores) en función de la fase y el desplazamiento.



Onda 1 y 3 tienen muchos casos, intensos de poco desplazamiento centrados en una fase particular.

Campos de regresión

Hace tiempo que quiero encontrar una buena forma de testear significancia en campos de regressión. El problema es que al hacer regresiones y testearlas punto a punto, uno termina haciendo múltiples comparaciones y no teniendo en cuenta la correlación espacial. Esto potencialmente crea falsos positivos a una tasa mayor que el alfa nominal elejido.

DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

Si X el el campo espaciotemporal con M celdas espaciales y x_i son las series temporales de cada celda, la regresión punto a punto con una serie temporal y implica hacer M regresiones lineales. Testear significancia de cada beta ingénuamente implica M tests estadísticos, lo cual genera problemas de multiplicidad.
Este paper de DeSole y Yang propone hacer una regresión lineal múltiple con M predictores.

$$\begin{aligned} y(t) &= \beta_1 x_1(t) + \epsilon_1(t) \\ y(t) &= \beta_2 x_2(t) + \epsilon_2(t) \\ &\dots \\ y(t) &= \beta_M x_M(t) + \epsilon_M(t) \end{aligned} \longrightarrow y(t) = \beta_1 x_1(t) + \beta_2 x_2(t) + \ldots + \beta_M x_M(t) + \epsilon(t)$$

Testear la significancia del campo de regresión de forma “global” implica testear que todos los betas son nulos, que en el caso de regresión múltiple, es testear el R^2 de la regresión.

DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

Problema:
Si la cantidad de celdas (M) es mayor que el largo de la serie temporal (casi siempre), entonces la regresión múltiple está sobredeterminada y no se puede hacer.
Solución:
Hay que reducir la dimensionalidad del problema. El candidato obvio es componentes principales.

$$X = UDV^t$$

Seleccionando k compoentes, se puede hacer la regresión:

$$y(t) = \beta_1 u_1(t) + \beta_2 u_2(t) + \ldots + \beta_K u_K(t) + \epsilon(t)$$

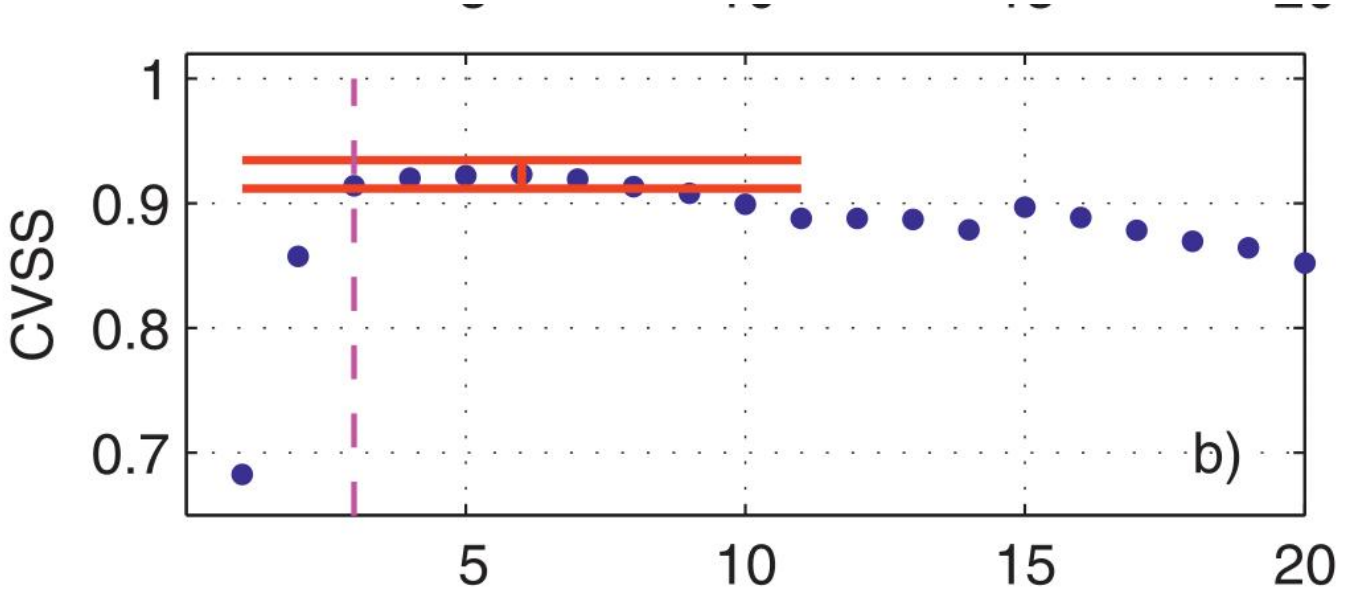
Y el campo de regresión se obtiene haciendo: BV^t

DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

Problema: (el de siempre)
¿Cómo elegir qué componentes usar?
Solución:
¯_(ツ)_/¯

DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

DeSole y Yang proponen quedarse con las primeras K componentes seleccionando el mínimo K que mazimiza el R^2 del ajuste.

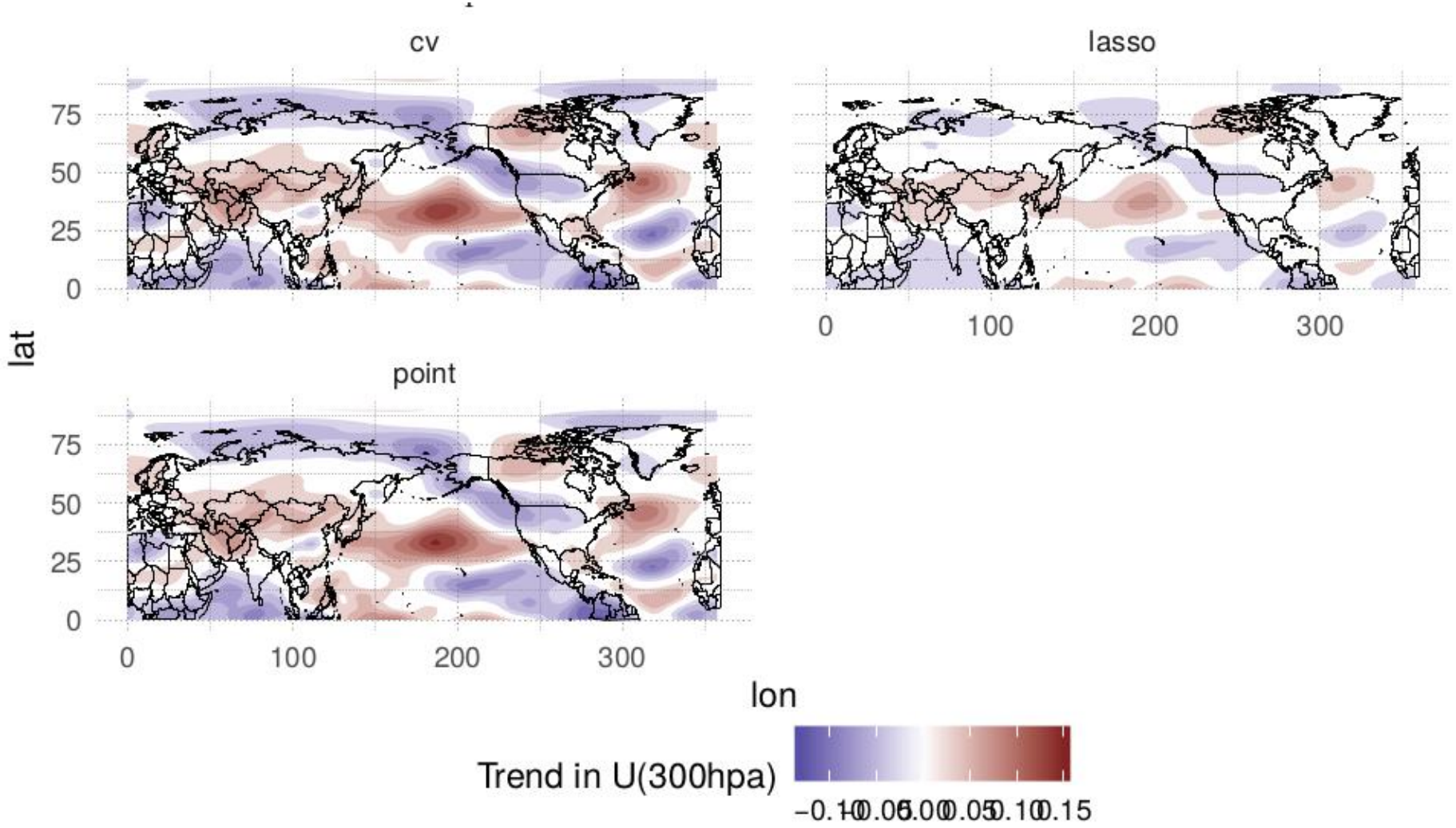


A mí se me ocurrió probar con regresión regularizada, que agrega una penalización a los coeficientes altos.

Alternativamente se puede proponer toda clase de metodologías de la literatura de selección de modelos. BIC, Akaike, etc..

DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

Ejemplo (replica el paper): tendencia lineal del viento zonal medio de DEF en 300hPa.



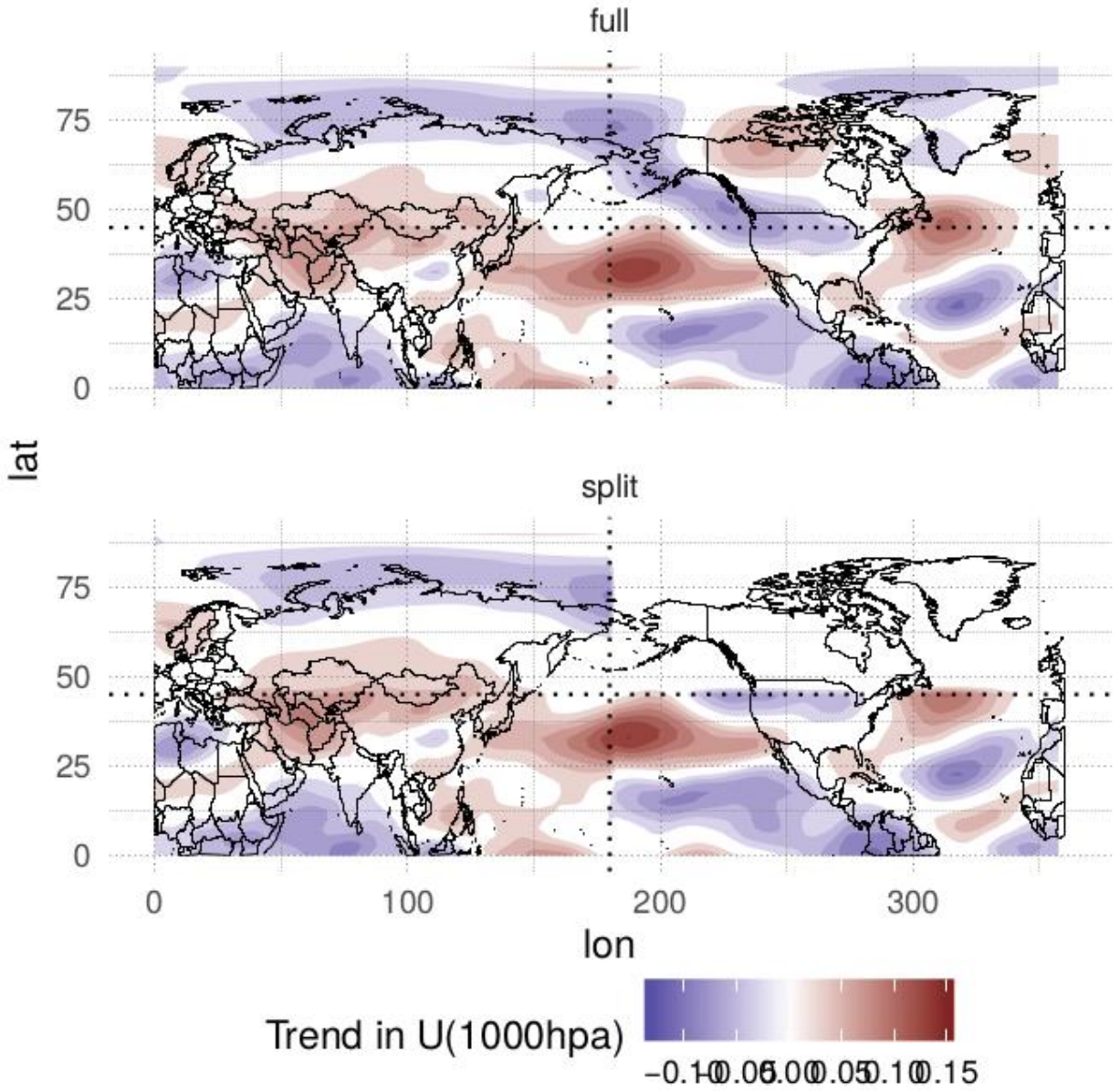
DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

Limitaciones:
Dependencia con el dominio.

Como el método pasa por componentes principales, está la siempre molesta dependencia con el dominio.

Arriba está el campo de regresión aplicando el método a todo el hemisferio norte. Abajo están los campos combinados aplicando el método a cada cuadrante por separado.

Hay diferencias en todo, pero el cuadrante superior derecho es el peor. Haciéndolo por cuadrantes, la regresión da casi nula. Esto es un problema, pero también indica que la regresion en esa zona no es estadísticamente significativa.



DelSole, T., & Yang, X. (2011). Field Significance of Regression Patterns. Journal of Climate, 24(19), 5094–5107.

Conclusiones

1. El método LASSO para reducir la dimensionalidad del problema no da buenos resultados. El p-valor obtenido es totalmente inválido y la candiad de coeficientes no nulos es muy sensible a la cantiad de compoentes principales que se permite entrar en la regresión. Además suele dar valores subestimados en la regresión (aunque eso es por diseño).
2. Hay algunos problemas con la elección del dominio que cambian el valor de la regresión. Sin embargo, estos cambios son informativos, ya que la variación es grande donde la señal es pequeña.
3. El p.valor conseguido es informativo!
4. Tiene una gran limitación: no acepta valores faltantes! Se puede usar DINEOF para rellenar usando EOF y tener algo consistente.

Localización de la onda

En la tesis de licenciatura hice una prueba muy preliminar de usar wavelets para localizar la onda 3 en latitud. Esto es una extensión de eso.

La idea que está detrás de la “localización” de una onda es la modulación de su amplitud. Es decir, que no hay una sola amplitud para todo el círculo de latitud, sino que la amplitud varía según la longitud. Un poco más estrictamente eso significa que si x es la variable que voy a descomponer en distintas ondas, lo que tengo en mente es:

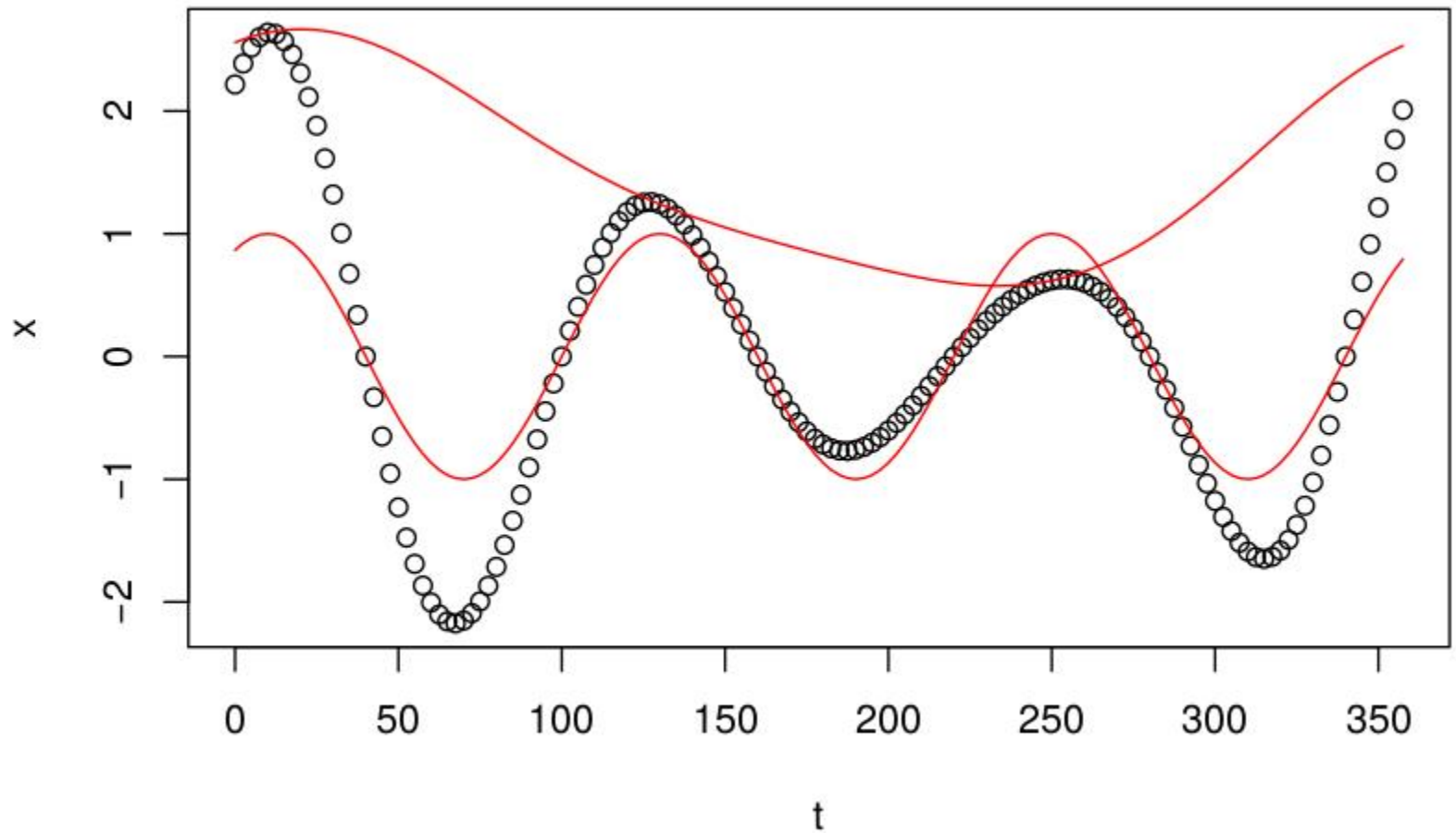
$$x = A(\lambda)\cos(k(\lambda - \phi_k))$$

Pero si el espectro de A incluye frecuencias mayores a k, ¿tiene sentido decir que está modulando la amplitud? Me parece razonable restringir la forma de A a suma de sinusoides con número de onda mayor a k. De manera que la ecuación anterior queda:

$$\tilde{x}_k = \sum_{k<k} [A_j \cos(j\lambda) + B_j \sin(j\lambda)]\cos(k(\lambda - \phi_k))$$

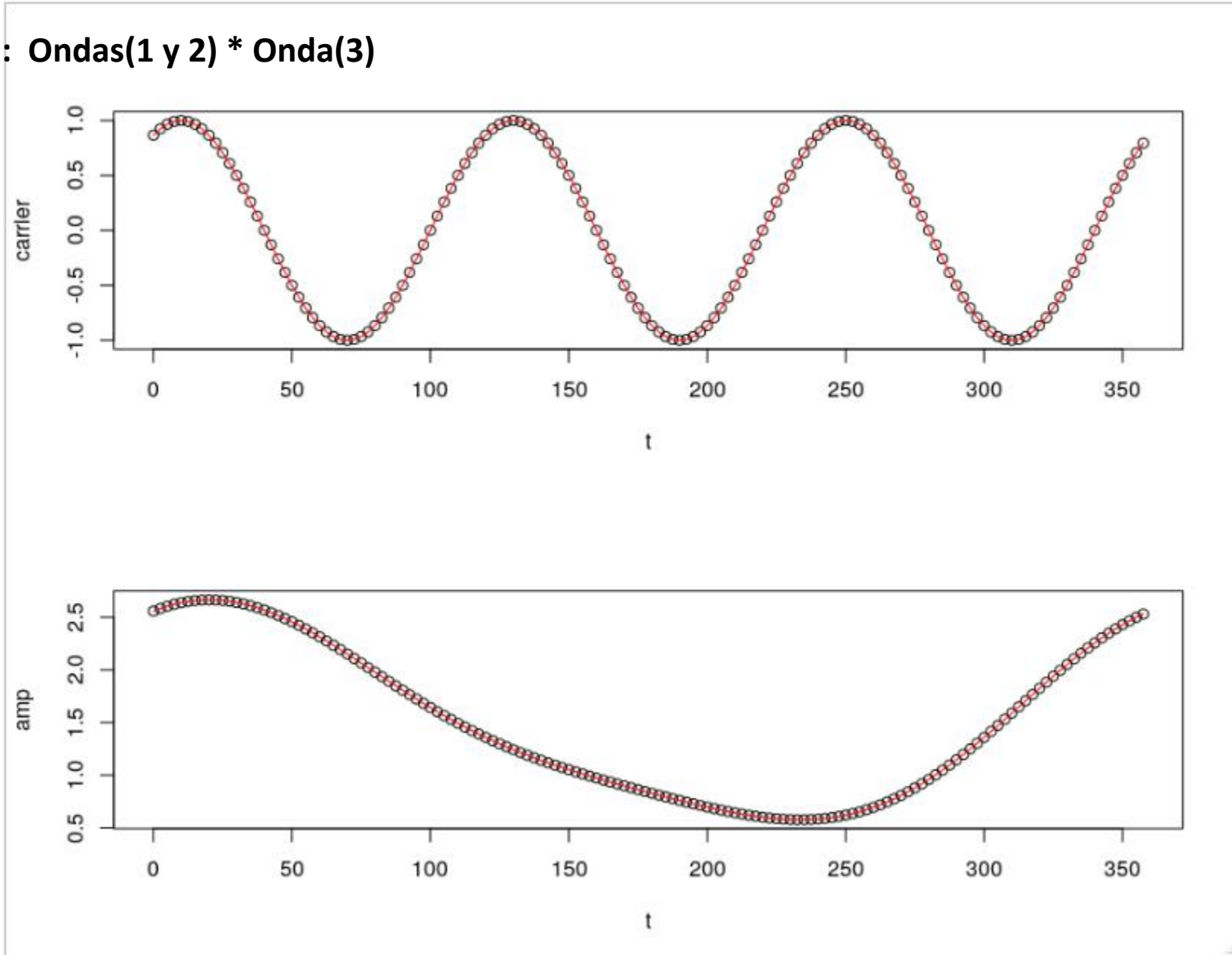
Cabe notar que como hay una multiplicación de senos y cosenos, el espectro de esta señal va a tener números de onda k + j y k - j.
De eso surge una primera limitación de este método: La modulación de una onda me “ensucia” el espectro con ondas maś largas y más cortas. Es imposible distinguir esas ondas de las ondas “independientes”.

Ejemplo ideal: Ondas(1 y 2) * Onda(3)



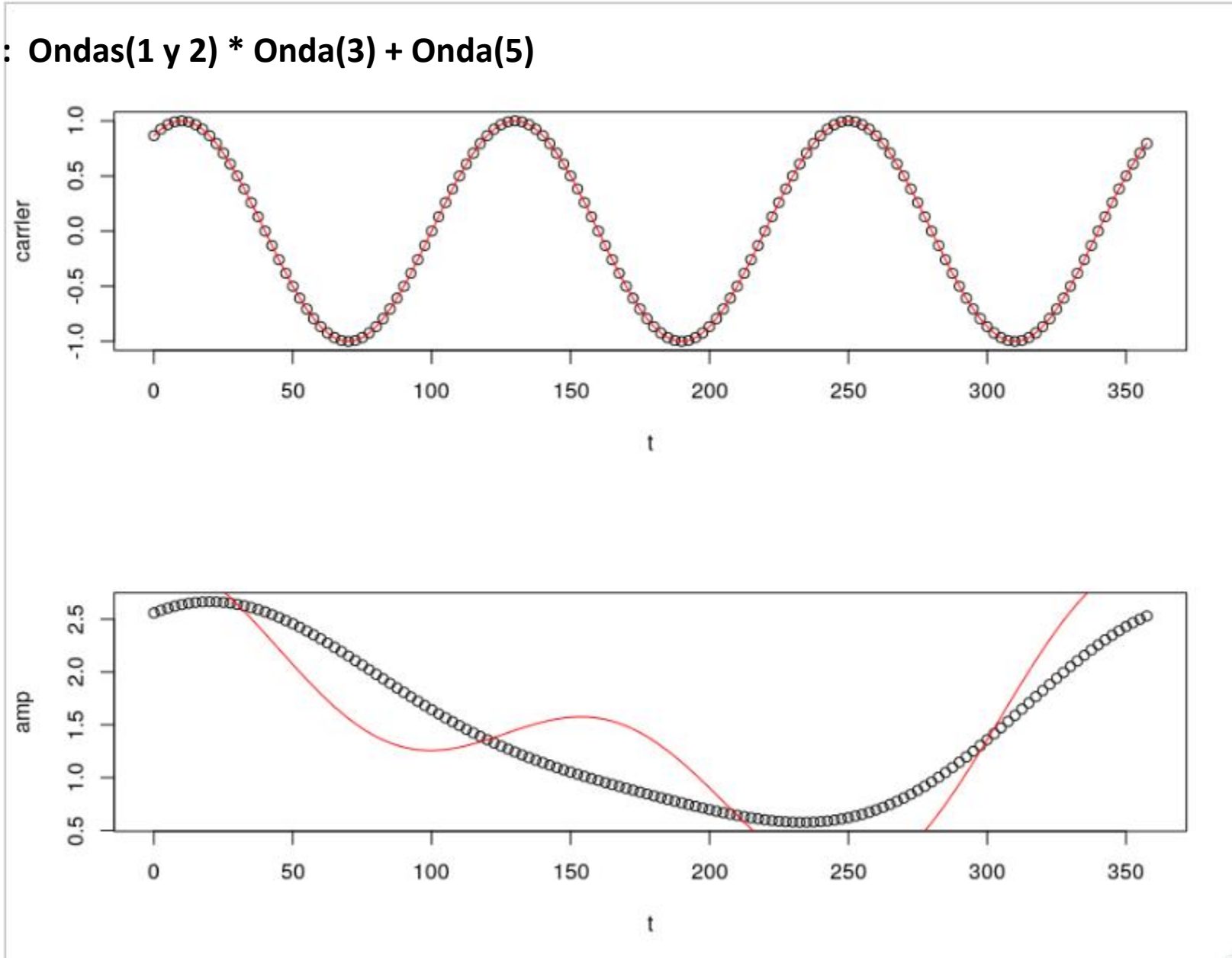
La señal en puntos es la onda 3 roja multiplicada por la envolvente roja (que tiene ondas 1 y 2)

Ejemplo ideal: Ondas(1 y 2) * Onda(3)



El método de demodulación recupera la onda 3 original y la envolvente .

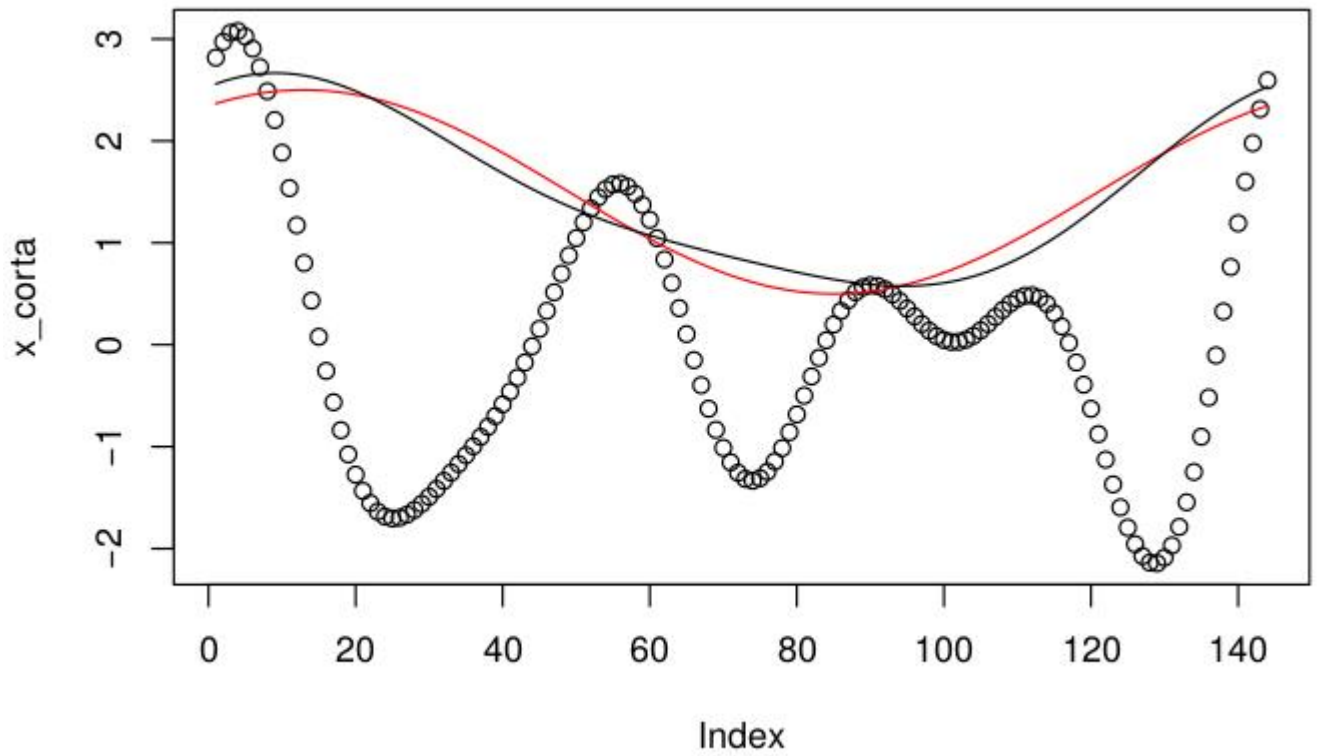
Ejemplo ideal: Ondas(1 y 2) * Onda(3) + Onda(5)



Si la señal tiene una más corta “individual”, ya no funciona bien.

Cómo reducir el “ruido”:

Hay que meter suposiciones más fuertes sobre A. Por ejemplo, que la única onda que modula es la 1. Esto equivale a modelar únicamente que la onda 3, por ejemplo, es más intensa de un lado que del otro del hemisferio.



Este es el caso anterior. Al limitarse a la onda 1, se elimina el efecto de la onda 5 sumada, pero se pierde el efecto de la onda 2 en la modulación.

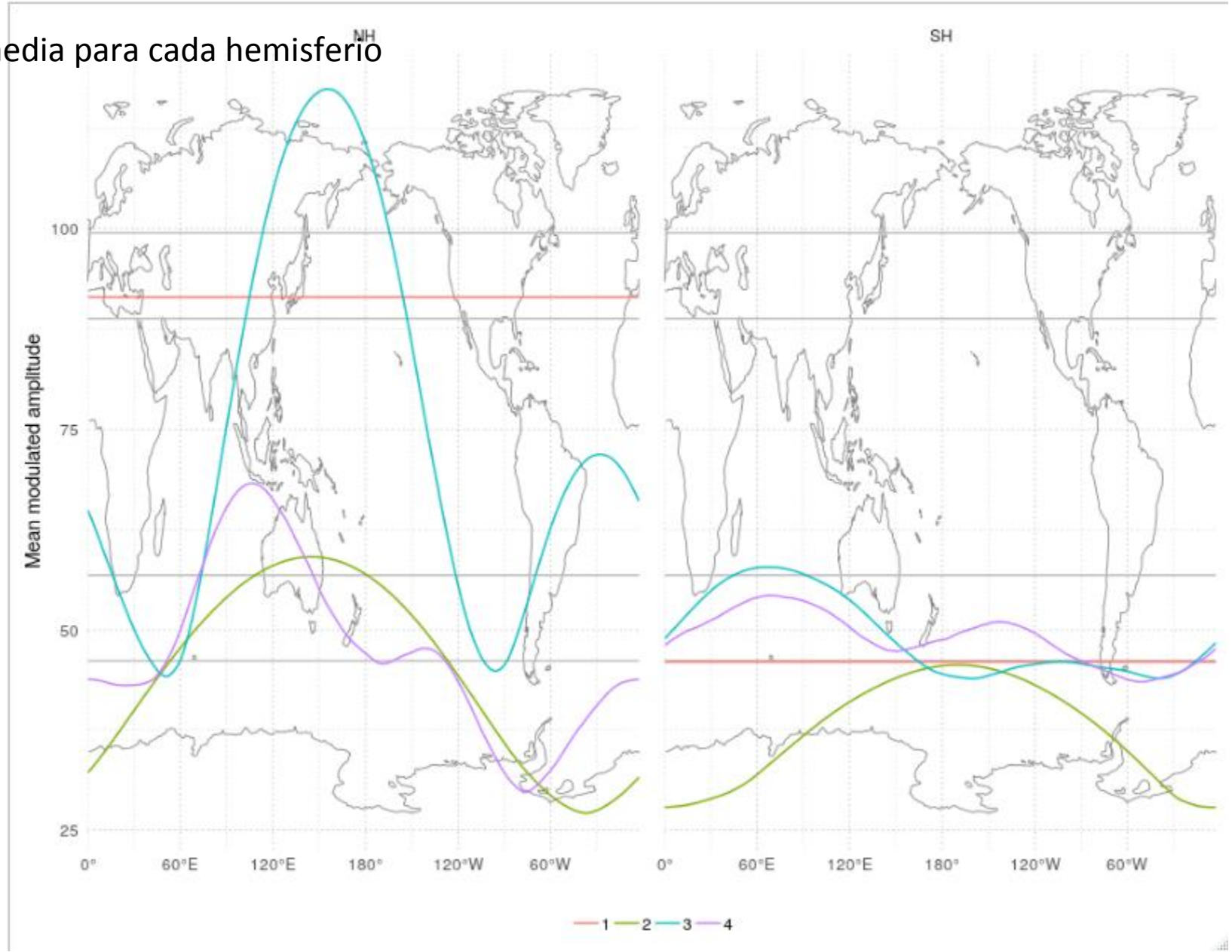
Lo que estoy haciendo, entonces, es calcular la envolvente de la señal, teniendo en cuenta sólo las ondas k+1, k y k-1.

Pero queda todavía una ambigüedad, porque a veces queda mejor la onda k por la envolvente y otras la k -1 por la envolvente.

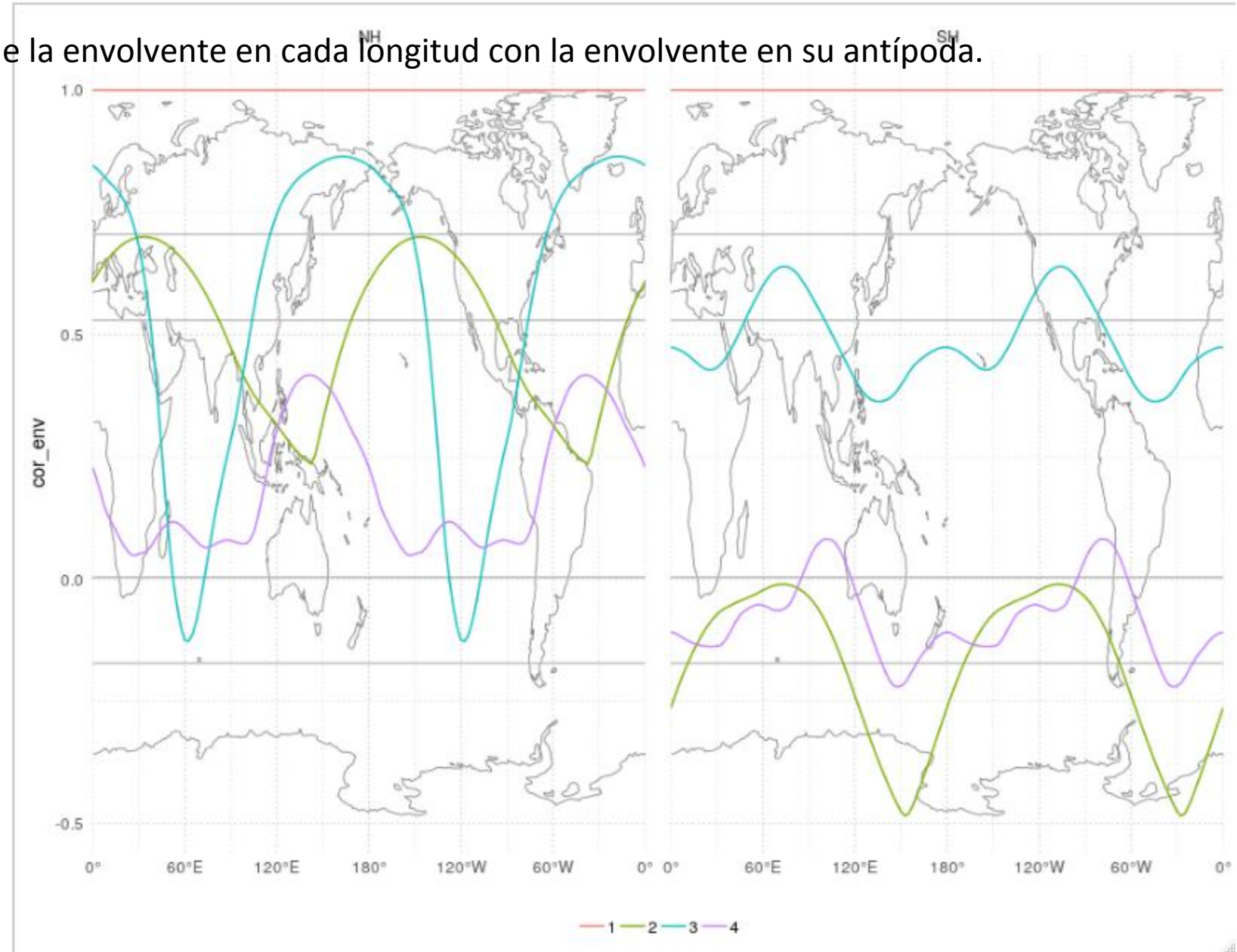
Lo que tomé entonces fue decidir que la onda “base” (k) es la onda con mayor R^2. Luego, calculo la envolvente de las ondas 1 a k.

Este es el caso anterior. Al limitarse a la onda 1, se elimina el efecto de la onda 5 sumada, pero se pierde el efecto de la onda 2 en la modulación.

Envolvente media para cada hemisferio

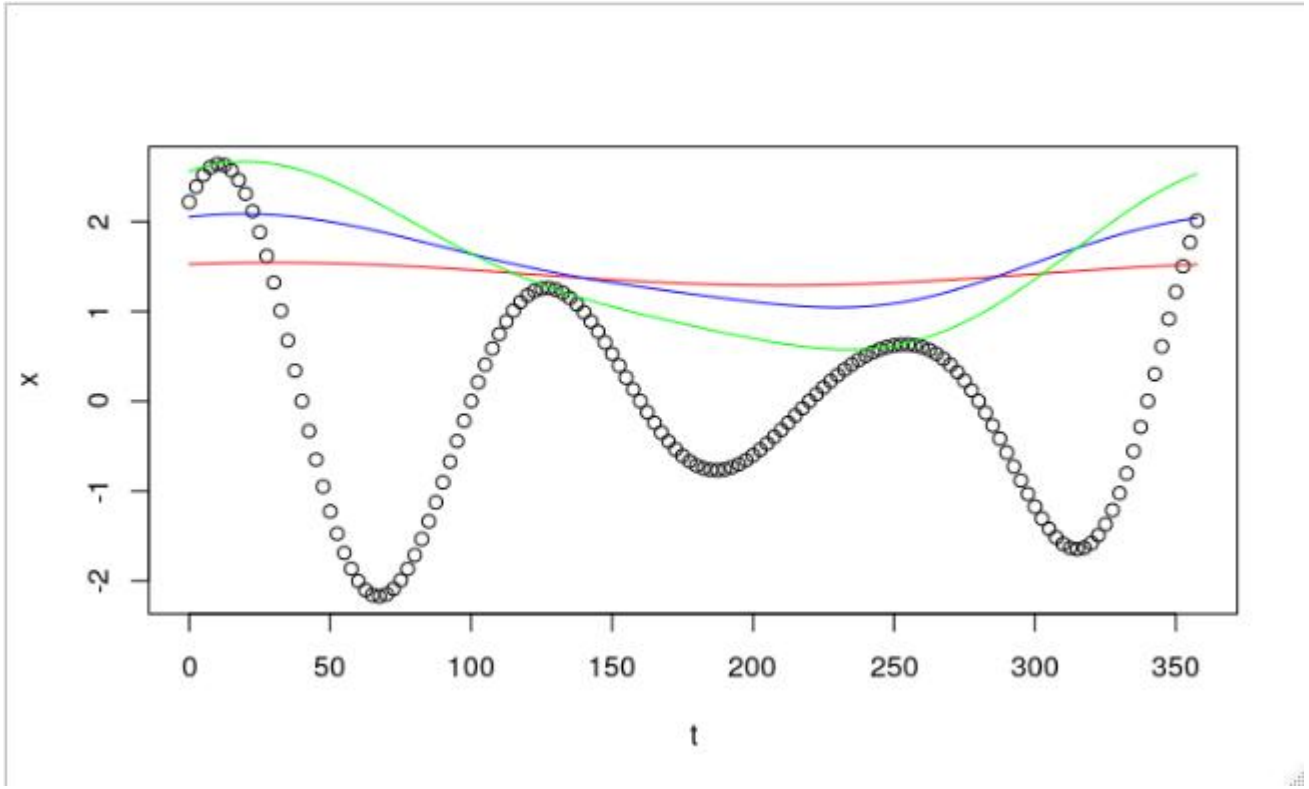


Correlación de la envolvente en cada longitud con la envolvente en su antípoda.



¿Cómo se compara esto con wavelets?

No lo tengo bien graficado, pero lo que puedo decir es que wavelets es mucho peor localizando la onda. La envolvente de wavelets es mucho más “homogénea” y con muy poca variabilidad zonal.



Verde: Envolvente real
Azul: Envolvente usando el método anterior
Rojo: Envolvente usando wavelets.