

Field Significance of Regression Patterns

Elio

Esto es un intento de implementar y entender el método propuesto por (DelSole and Yang 2011) para calcular la significancia de los patrones de regresión.

Idea básica

La idea es que dado una serie de campos bidimensionales a lo largo del tiempo uno puede armar un campo de regresiones a partir de una serie temporal. Si el campo tiene M celdas, entonces hay que calcular M regresiones:

$$y(t) = \beta_1 x_1(t) + \epsilon_1(t) \\ y(t) = \beta_2 x_2(t) + \epsilon_2(t) \\ \vdots \\ y(t) = \beta_M x_M(t) + \epsilon_M(t)$$

Donde $x_i(t)$ representa el valor de la variable espacial en la celda i -ésima y el campo de regresión son los coeficientes $\beta_1, \beta_2, \dots, \beta_M$. Testear la significancia de este campo no es trivial. En general lo que se hace es testear cada β_i por separado como si fuera independiente y obtener un “campo de significancia”. Pero eso en realidad no es del todo válido porque ignora la correlación espacial entre los coeficientes y, más importante aún, la multiplicidad de tests.

(DelSole and Yang 2011) propone cambiar el problema. En vez de hacer M regresiones simples, hacer una regresión múltiple:

$$y(t) = \beta_1 x_1(t) + \beta_2 x_2(t) + \dots + \beta_M x_M(t) + \epsilon(t)$$

para la cual se puede testear la $H_0 := B_i = 0 \forall i = 1, 2, \dots, M$ sin más dificultad. Esto elimina el problema de multiplicidad y la correlación espacial. Simple, ¿no? Listo..

No tan rápido.

Selección del modelo

Esta regresión funciona si la cantidad de predictores (M) es menor que la cantidad de datos usados para realizar el ajuste (N), pero en la mayoría de los casos hay más celdas que datos. Por ejemplo, un campo global de 2.5° de resolución se tiene $144 \times 72 = 10368$ celdas. Si queremos hacer un mapa de regresión con la ecuación anterior usando datos mensuales serían necesarios más de 864 años de datos. Imposible.

El truquito ahora está en reconocer que la correlación espacial implica que en un campo global no hay 10368 predictores independientes; existe mucha redundancia de información. Una forma de aprovechar esto es hacer la regresión en el espacio de las componentes principales usando sólo algunas (y siempre menos que N) y luego reconstruir el campo conseguido. Si la matriz de datos es X (donde cada columna es un $x_i(t)$), hacemos

$$X = UDV^t$$

Y pasamos a hacer la regresión

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

Donde u_i son las columnas de U . Con $K < N$ nos aseguramos que la regresión ande bien. Luego, el campo de regresión es

$$BV^t$$

(hay algunas constantes de normalización dando vuelta que no son demasiado importantes a la teoría –pero sí a la práctica!)

Simple, ¿no? Listo..

No tan rápido.

¿Cuántas y cuáles componentes principales elegir? Es el problema eterno. El paper propone seleccionar las primeras K componentes principales usando validación cruzada para juzgar el “mejor K ”. El procedimiento es el siguiente:

Para $k = 1$ se ajusta el modelo N veces dejando de lado una observación por vez y computando la diferencia entre el y observado y el modelado. Se consiguen N errores y de estos se computa el MSE y se asume que éste tiene un desvío estándar de MSE/\sqrt{N} . Se repite el procedimiento para todos los k . El resultado es una serie de $MSE(k)$. Se elige el mayor k para el cual el MSE esté dentro del intervalo de 1 desvío estándar del menor MSE observado.

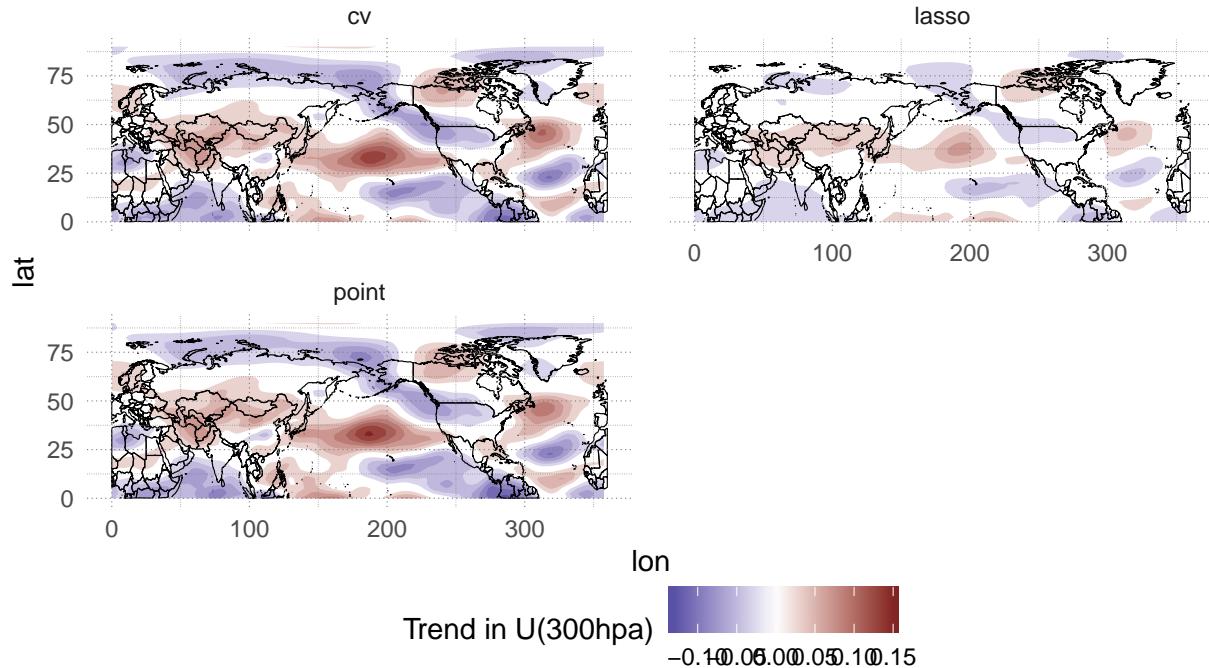
De forma general, esto es un problema de *feature selection* que tiene otras soluciones. Una alternativa es usar [LASSO](#) o [ridge regression](#), que penaliza los coeficientes “grandes”.

Ejemplos

Viento zonal

Voy a usar datos de viento zonal medio de DEF en 300hPa en el hemisferio norte para replicar lo que hicieron los autores del paper.

Una aclaración importante es que los autores del paper usan leave-one-out crossvalidation, pero eso es **eterno** para una cantidad de datos medianamente grande, así que yo implementé k-fold crossvalidation con $k = 10$ por default.



Los campos de regresión son prácticamente iguales. La naturaleza de LASSO, que penaliza coeficientes grandes, hace que los valores absolutos sean menores. Podemos ver cuándos EOFs se usaron en la regresión, el r^2 del fit y su p-valor siguiendo la metodología del paper

	r2	f.statistic	p.value	non_zero
cv	0.88	14.92	8.00000e-16	20
lasso	0.62	6.22	3.83875e-10	13

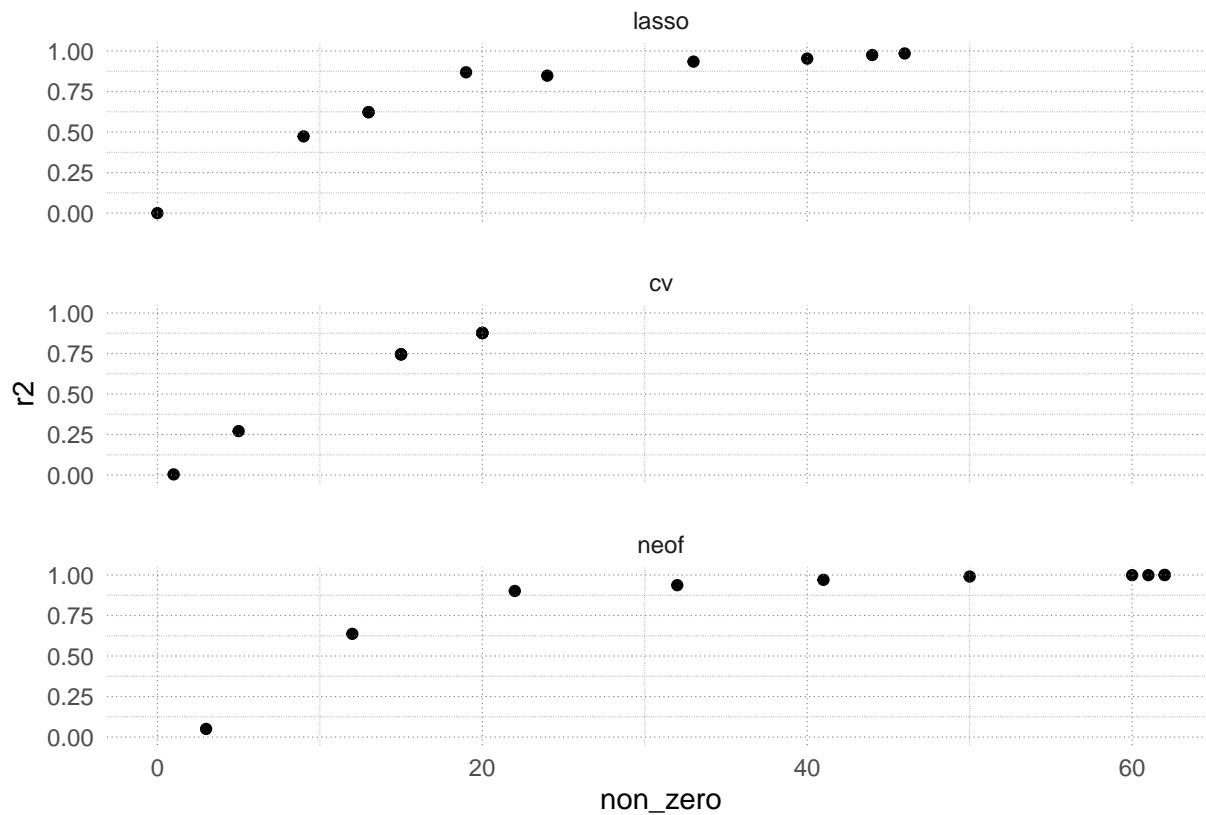
Aunque los patrones son similares, usan distinta cantidad de EOF. La estimación del p-valor no sé si es válida para LASSO, ya que no es cuadrados mínimos, así que la distribución del r^2 posiblemente no sea la que dicen en el paper. Pero para ser justos, tampoco sé si con la validación cruzada sigue valiendo.

Un detalle para el caso de LASSO es que los EOFs que usa varían muchísimo si cambio la cantidad de EOFs que permito entrar a la regresión desde un principio

max_eof	non_zero_lasso	non_zero_cv	r2_lasso	r2_cv
3	0	1	0.00	0.00
12	9	5	0.47	0.27
22	19	15	0.87	0.74
32	24	15	0.85	0.74
41	33	15	0.93	0.74
50	46	20	0.98	0.88
60	44	20	0.98	0.88
61	40	20	0.95	0.88
62	13	20	0.62	0.88
63	13	20	0.62	0.88

En particular, se ve que la cantidad de coeficientes no nulos aumenta a medida que se ponen más eofs pero luego disminuye mucho cuando se ponen cerca del máximo. La crossvalidación, en cambio, se queda estable y no es sensible a ese problema (al menos para este ejemplo!)

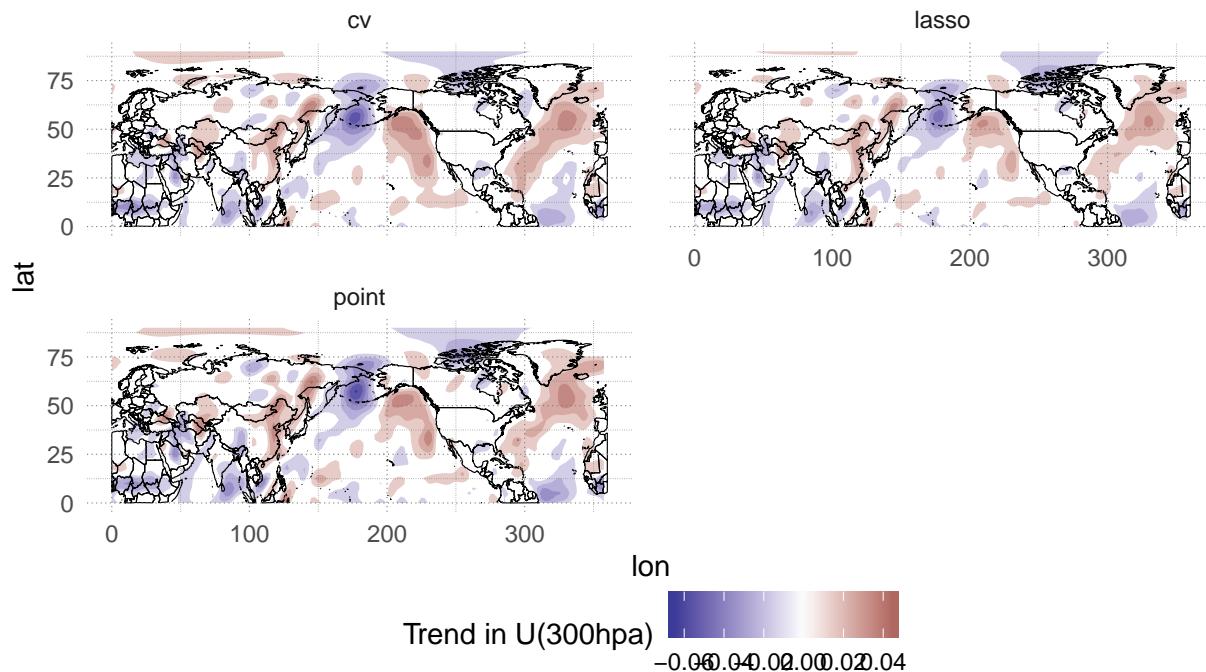
El r^2 también tiene un comportamiento similar. Comparando el r^2 en función de la cantidad de coeficientes no nulos para el método LASSO, crossvalidación y el “ingénuo” se ve que el método de crossvalidación deja de incluir EOFs en el “codo” de la curva, mientras que LASSO los sigue incluyendo.



¿Por qué luego baja cuando se incluyen todos los EOFs? No sé. Es posible que tenga que ver con el algoritmo de LASSO, que se comporta diferente según la cantidad de variables en la regresión. Misterio.

Viento meridional en 1000hPa

Para probar con otros datos.

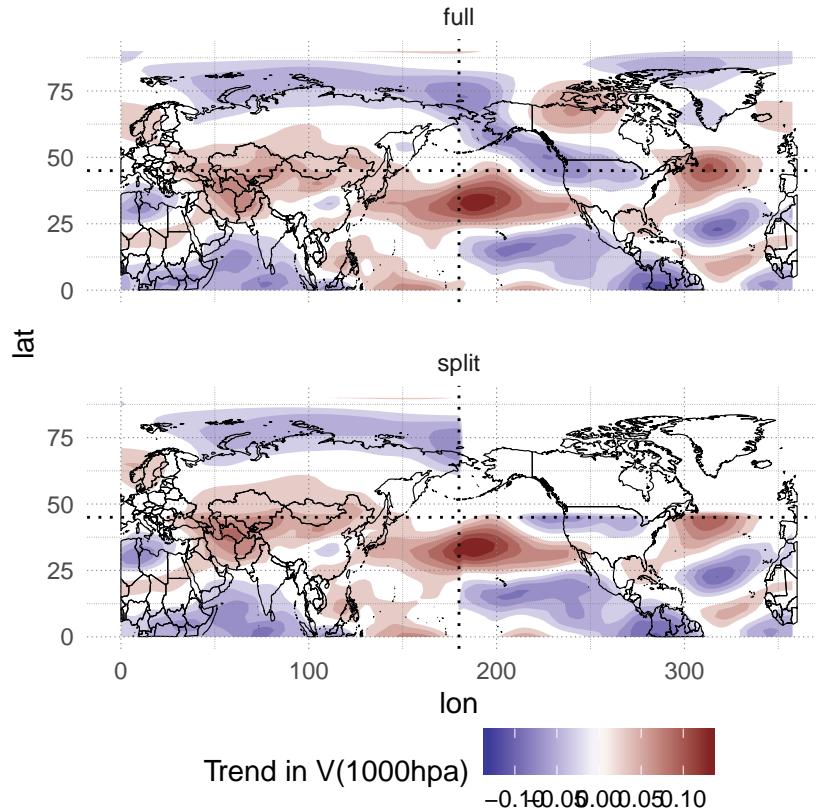


En estos datos, ambos métodos dan buenos resultados caracterizando el campo de regresión. LASSO tiene los mismos problemas que en el caso anterior.

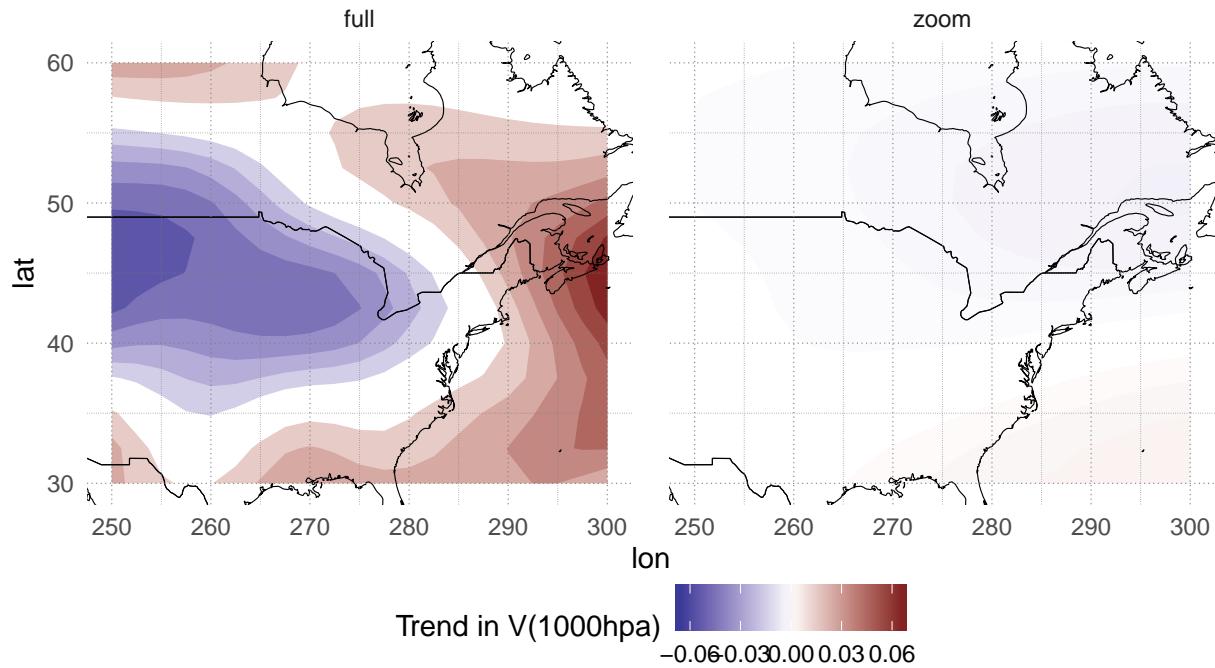
Problemas

Dependencia con dominio

Un problema de los EOFs es que pueden depender mucho del dominio. Por lo tanto, es esperable que los mapas de regresión también dependan del dominio si se usan EOFs para generarlos. Por ejemplo, estos son los mapas de regresión para U en 300hPa calculados usando todo el dominio o sólo un hemisferio por vez.



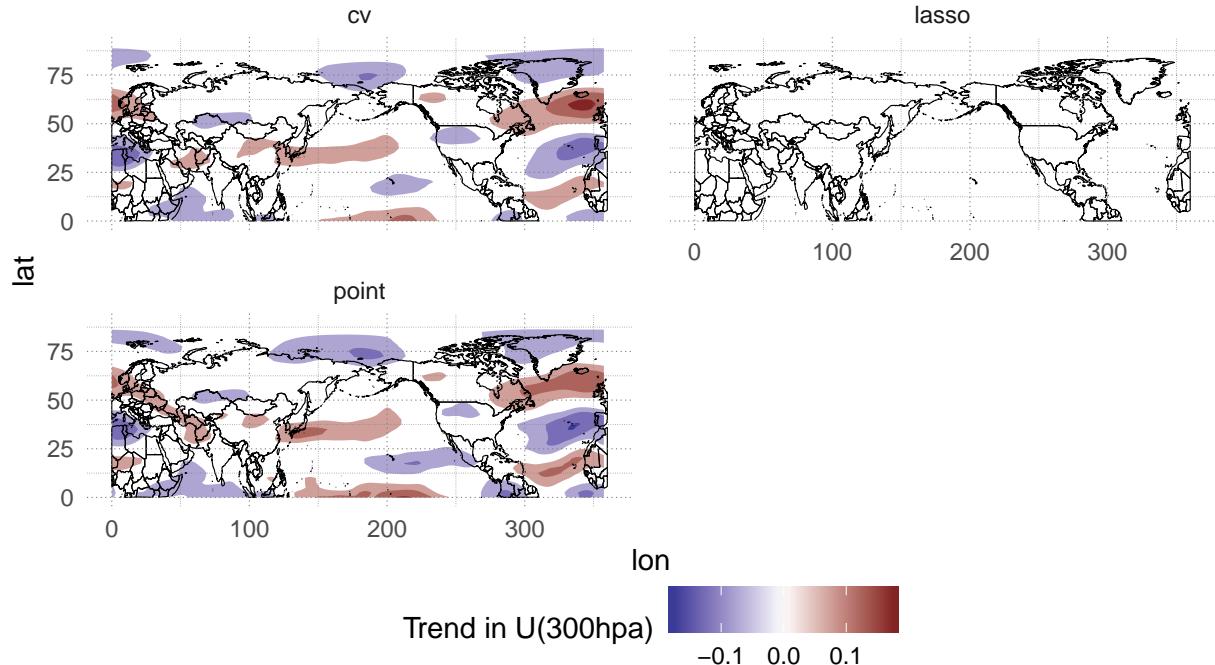
En algunas partes hay diferencias, aunque en otras no tanto. El cuadrante noroeste perdió toda su magnitud. Si hacemos un “zoom”, el resultado de calcular la regresión en el dominio recortado o recordar la regresión global puede ser muy distinto:



La diferencia en la magnitud es enorme, pero también la distribución. El resultado del zoom no es estadísticamente significativo.

Dependencia con la longitud de la serie

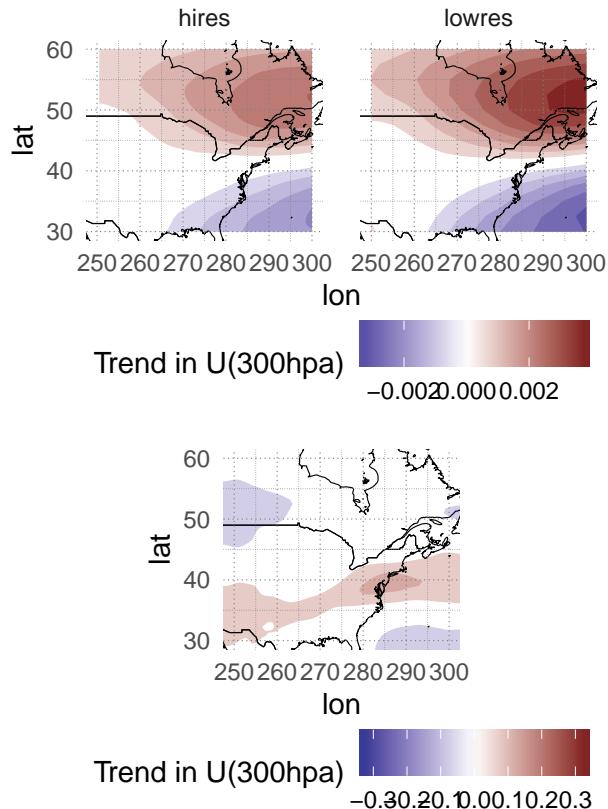
¿Qué pasa si saco datos? Digamos que tengo un tercio de los datos (21 años).



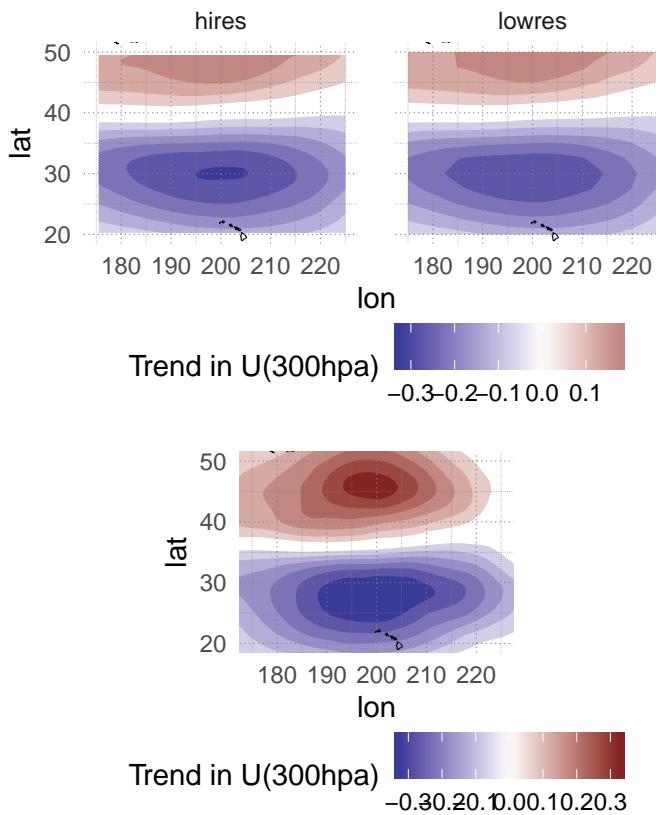
Las estimaciones cambian, obviamente, pero se ve que LASSO prácticamente no encuentra regresiones importantes. Hay un solo EOF con coeficiente no nulo para LASSO, pero el p.valor calculado por la fórmula estándar da 0.0058. O sea, bastante significativo. El p.valor de crosscorrelación es 0.0388; no muy significativo.

Dependencia con la resolución

¿Depende de la resolución de los datos? Voy a usar datos de ERA Interim con 2.5x2.5 para la baja resolución y con 0.75x0.75 para la alta resolución.



Las distintas resoluciones dan algo ligeramente distinto, pero no mucho. El tema es que analizando los coeficientes, sólo la primera componente principal da no nula y no se parece en nada a la tendencia punto a punto. Además notar la enorme diferencia en la magnitud del efecto. Lo que está pasando es que el ajuste no es significativo ($p\text{valor} = 1$ para ambas resoluciones). Si pasamos a otra región, la cosa es distinta:



En esta región del pacífico donde las tendencias son más intensas, el recorte no cambia mucho el patrón de regresión. La regresion es estadísticamente significativa en ambas resoluciones. La intensidad del cambio es similar.

Conclusiones

1. El método LASSO para reducir la dimensionalidad del problema no da buenos resultados. El p-valor obtenido es totalmente inválido y la cantidad de coeficientes no nulos es muy sensible a la cantidad de componentes 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 llenar usando EOF y tener algo consistente.

Bibliografía

DelSole, Timothy, and Xiaosong Yang. 2011. "Field significance of regression patterns." *Journal of Climate* 24 (19): 5094–5107. <https://doi.org/10.1175/2011JCLI4105.1>.