Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Trabajo de Especialización

Modelo de pronóstico estadístico de precipitación estacional para el verano aplicado a la región Pampeana.

Autor

C.C. Alfredo Luis Rolla

Indice General

Modelo de pronóstico estadístico de precipitación estacional para el verano aplicado a la región Pampeana.

Resumen

Se estudió la variabilidad interanual de la lluvia de verano en la región Pampeana para intentar predecir la precipitación estacional utilizando modelos de regresión lineal múltiple. Este trabajo utilizó las precipitaciones observadas de verano (diciembre, enero y febrero) como predictando en la región Pampeana de Argentina, y como variables predictoras series de variables atmosféricas y oceánicas globales del mes anterior al verano (noviembre). Se emplearon técnicas de agrupamiento para definir regiones similares de precipitación estacional de verano. Las principales variables predictoras globales utilizadas fueron la altura geopotencial (presión atmosférica), cantidad total de agua en la columna atmosférica, viento en capas bajas de la atmósfera (850hPa) y la temperatura de la superficie del mar. El análisis estadístico de las mismas permitió definir las series de variables predictoras a ser utilizadas en cada región para la construcción de modelos de regresión lineal múltiple regionales. Finalmente se aplicaron los modelos regionales a una situación real de pronóstico.

1. Introducción.

Las adversidades climáticas que enfrenta el sector agrícola durante el proceso de producción generan un alto grado de incertidumbre sobre el resultado de la actividad, que lleva alto nivel de riesgo asociado a las unidades productivas. Dada la diversidad de climas y suelos en Argentina, todos los agricultores enfrentan el riesgo de pérdidas debido a factores climáticos, ya sea por la sequía, heladas, granizo, exceso de agua, fuertes vientos o inundaciones, entre otras adversidades.

Los resultados en la producción agrícola dependen de los pronósticos y las estrategias adoptadas, que generalmente tienen en cuenta un comportamiento normal de las variables climáticas. Estas estrategias son desarrolladas por el propio agricultor en su región de interés y fundamentalmente tienden a reducir la vulnerabilidad frente a condiciones climáticas adversas. Un ejemplo de estas estrategias es la protección de activa de cultivos, por ejemplo, es posible mencionar la aplicación de riego por aspersión como un método para reducir el impacto de las heladas o aplicación de riego suplementario para reducir el déficit de agua o la evaluación de las condiciones climáticas ideales para la siembra, la fertilización o el momento de la utilización de maquinaria para la cosecha de cultivos (por exceso de precipitación) entre otros.

Por lo tanto, tener un buen conocimiento de la variabilidad de la precipitación regional es importante para los tomadores de decisión, particularmente para la implementación de estas estrategias sobre los cultivos.

Estudios previos han detectado un desplazamiento hacia el oeste de las isoyetas de precipitación acumulada anual en la Argentina subtropical desde mediados del siglo XX (Liebmann et al., 2004, Barros et al., 2008, Saurral et al., 2016), entre otros.

Sin embargo, los datos de las últimas décadas muestran cierta evidencia de cambio en este comportamiento en regiones específicas, lo que provoca pérdidas económicas y mayores problemas sociales. Por ejemplo, se han observado tendencias negativas de precipitación en algunas áreas localizadas en la región del Chaco argentino (González et al., 2012a). El conocimiento de la variabilidad interanual de la precipitación es

importante para detectar el efecto producido por los forzantes de gran escala sobre la precipitación, especialmente en América del Sur.

Es sabido también que El Niño-Oscilación del Sur (ENSO) tiene una gran influencia sobre la precipitación en el Sudeste de América del Sur (SESA) (Ropelewski y Halpert, 1987, Kiladis y Díaz, 1989, Compagnucci y Vargas, 1998, Grimm y otros, 2002, Vera et al. 2004, Barreiro, 2010, Garbarini et al., 2016, entre otros). Ashok et al. (2007) definieron un tipo especial de El Niño, que llamaron El Niño Modoki (EMI), donde la zona de interés de calentamiento se limita al Pacífico Central tropical, mientras que al este y al oeste de esta zona central se encuentran zonas de enfriamiento. Estos eventos no producen exactamente los mismos efectos sobre la precipitación que los registrados por los eventos tradicionales de El Niño.

Otro forzante climático global relacionado con la precipitación y la temperatura de la superficie del mar (SST) es el Dipolo del Océano Índico (IOD) (Saji et al., 1999), cuya fase positiva se define por el calentamiento del Océano Índico Sudoccidental y el enfriamiento del Océano Índico Noreste. Estas áreas son las que definen el índice IOD que describe este proceso. Chan et al. (2008) encontraron que en América del Sur la fase positiva de la IOD se manifiesta como un dipolo de anomalías de precipitación, con aumentos en la cuenca del Plata y disminuciones en la región del centro de Brasil.

Otro factor asociado a los vientos que influye en la región es la Oscilación Antártica (AAO) (Thompson y Wallace, 2000), cuya fase positiva está definida por anomalías de presión negativa alrededor del Polo Sur y anomalías subpolares bajas y anomalías positivas de presión formando un anillo en el área del alta subtropical. El índice se define como la diferencia de presión normalizada entre 40 ° S y 70 ° S (Nan y Li, 2003). La fase positiva está asociada con una intensificación del viento zonal y por lo tanto con un intercambio menor entre latitudes altas y medias. Hay varios trabajos que investigan los efectos de AAO en el clima de América del Sur. Reboita et al. (2009) encontraron que la función de frontogénesis es intensa durante la fase negativa de AAO y la trayectoria de la actividad ciclónica se desplaza hacia el sur durante la fase positiva de la AAO.

González y Vera (2010), González et al. (2010), González y Domínguez (2012) y González (2015) encontraron una relación significativa entre las fases negativas de la AAO con las lluvias de invierno en el área de Comahue en los Andes, al noroeste patagónico.

El dipolo del Atlántico Sur (SAODI) (Nnamchi et al., 2011) es una oscilación que se ha definido en el Océano Atlántico. Su fase positiva se define como el calentamiento del noreste (costa de África) y el enfriamiento del suroeste del océano tropical (costa de Brasil). Este efecto está relacionado con la posición e intensidad del anticiclón semipermanente del Atlántico Sur y genera anomalías en las advecciones de humedad que se manifiestan en anomalías de precipitación.

En particular, el área de estudio seleccionada para este trabajo es la región Pampeana, la principal área de producción agrícola de Argentina. Desde el punto de vista económico, tres cuartas partes del valor total de la producción agrícola corresponden a esta zona, que abarca 5 millones de hectáreas. Sólo las provincias de Buenos Aires, Santa Fe y Córdoba generan el 70% de la producción agrícola del país.

2. Datos y Metodología.

Se utilizaron datos diarios de precipitación de 37 estaciones meteorológicas con registros completos y que cubrieran el área de estudio (Figuras 1a y 1b) pertenecientes a la red de medición del Servicio Meteorológico Nacional de Argentina (SMN) en la región Pampeana para el período 1979-2016. Los datos fueron consistentes y solo se usaron aquellos con alta calidad. Los datos faltantes no excedieron el 1% de los datos diarios totales. En base a las series de observaciones diarias de precipitaciones de las estaciones meteorológicas se generaron las series acumuladas estacionales (DEF de diciembre a febrero, verano; MAM de marzo a mayo, otoño; JJA de junio a agosto, invierno; SON de septiembre a noviembre, primavera). Para este estudio solo se utilizó la precipitación estacional de verano (DEF). Se realizó un análisis exploratorio de estas series de datos usando metodologías estadísticas.

Se definieron subregiones homogéneas de estaciones meteorológicas para generar los modelos de pronóstico de precipitación de verano. Para ello, se evaluaron distintos métodos de agrupamientos jerárquicos (Tan, 2006) del tipo aglomerativos utilizando un par de métricas de distancia que fueran coherentes con el problema a resolver como son la distancia euclídea y la correlación entre estaciones, además se probaron métodos de agrupamiento no jerárquicos como K-means (Tan, 2006; MacQueen,1967; Steinhaus 1956). Para verificar el agrupamiento se emplearon trabajos preexistentes realizado por expertos en el tema. Una metodología común utilizada en meteorología para realizar este tipo de agrupamientos es el método de Lund (1963) que se basa en un agrupamiento jerárquico usando una matriz de distancia basada en la correlación.

Para cada grupo se estudiaron los forzantes de la variabilidad interanual de la precipitación de verano mediante el cálculo de las correlaciones entre la precipitación de verano y las variables meteorológicas y oceánicas: de altura geopotencial (hgt) en 1000, 500 y 200 hPa, viento zonal (u) y meridional (v) en 850 hPa, agua precipitable (tcw) en la capa desde superficie hasta 700 hPa y temperatura de superficie del mar (sst) obtenidas de los reanálisis NCEP/NCAR (Kalnay y otros, 1996). El conjunto de datos de reanálisis NCEP / NCAR es un conjunto de datos grillado global que es actualizado continuamente (1948-presente) que representa el estado de la atmósfera, estos reanálisis se construyen incorporando observaciones y Pronósticos de modelos numéricos meteorológicos (NWP). Es un producto conjunto de "National Centers for Environmental Prediction" (NCEP) y el "National Center for Atmospheric Research" (NCAR). Se utilizaron estos datos en el mismo período que las observaciones (1979-2016), así seleccionado, porque se considera que a partir de 1979 la red de observaciones global incorporó información satelital, mejorando significativamente la calidad y veracidad de los reanálisis a nivel regional (hemisferio sur), que fueron usados como predictores en los modelos de regresión. Estos datasets/reanálisis tienen una resolución espacial de 250 km y en este caso una resolución temporal mensual, en particular el mes de noviembre.

Los mapas de correlaciones se calcularon entre las observaciones de precipitación de series de tiempo estacionales (DEF) versus las variables de gran escala provenientes de los reanálisis descriptos anteriormente (hgt, u, v, tcw, sst) del mes anterior al trimestre de verano (noviembre). Las correlaciones se consideraron significativas con el 95% de confianza cuando superaron el umbral de 0,35. El resultado de estas correlaciones fueron mapas globales de correlaciones.

Con los mapas globales de correlaciones construidos con las series regionales de observaciones estacionales medias de precipitación y los forzantes, se seleccionaron todas aquellas áreas o convexos con correlaciones significativas que tuvieran un significado físico asociado a efectos conocidos sobre la precipitación en la región respecto de la dinámica de la atmósfera global y su influencia en el hemisferio sur. Sobre estas regiones con correlaciones significativas se calcularon las series correspondientes al mes de noviembre de la variable considerada.

Luego de esta pre-selección de áreas y de predictores, se utilizó la metodología de LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) que es un método de análisis de regresión que realiza la selección y regularización de los predictores que permitiría mejorar la precisión de la predicción e interpretabilidad de los modelos estadísticos generados, descartando aquellos predictores que explican poca varianza de la precipitación estival para la construcción de los modelos de regresión lineal múltiple. Además, se consideraron predictores que fueran linealmente independientes entre sí para evitar el problema de la multicolinealidad entre predictores.

Finalmente, con estos predictores se construyeron modelos de regresión lineal múltiple para cada una de las áreas homogéneas para realizar el pronóstico de la precipitación de verano. Los modelos resultantes se validaron utilizando la metodología de validación cruzada (Stone,1974), en particular, la metodología de "Leave-one-out cross-validation" (LOOCV). Esta técnica se eligió debido a la poca cantidad de observaciones disponibles. En este método se utilizan todos los años menos uno para la construcción del modelo y el año restante para la predicción. El proceso fue repetido tantas veces como años se utilizaron en este estudio para la construcción de los modelos en particular se usó el periodo (1979-2015). Este proceso permitió al mismo tiempo validar el pronóstico y verificar la estabilidad de los modelos.

Para probar la eficiencia del modelo, se confeccionaron tablas de contingencia de la precipitación observada y la pronosticada, donde se separaron los casos equiprobables llamados, subnormales, normales y sobrenormales, refiriéndose cada uno al espectro completo de casos posibles desde los años más secos hasta los años más húmedos, respectivamente. Además, se calcularon los siguientes índices para evaluar el ajuste de los modelos: probabilidad de detección (**recall**), relación de falsa alarma (**FNR**) y porcentaje de aciertos (**accuracy**) y se compararon las funciones de probabilidad acumulada de la precipitación observada y estimada por los modelos usando el test de chi-square para evaluar el ajuste de los modelos.

El software utilizado para realizar este estudio fue lenguaje R y R-Studio (R Core Team, 2017) como interface de desarrollo.

Los principales paquetes utilizados fueron 'RMySQL' (Ooms et al, 2018), 'ncdf4' (Pierce et al, 2017), 'dplyr' (Wickham et al, 2017) para el manejo de datos.

Se utilizaron los paquetes 'rgdal', 'maptools' (Bivand et al, 2018) y 'raster' (Hijmans, 2017) para los calculos con datos georeferenciados.

Para clusterizar se uso el paquete 'cluster' (Rousseeuw et al, 2018) y 'ggdendro' (de Vries, 2016)

Para crear los modelos de regresion lineal multiple se uso 'glmnet' (Hastie Trevor, 2018) y 'fpp' (Hijmans, 2017).

Los graficos en general se hicieron usando 'ggplot2' (Wickham et al, 2017).

3. Resultados.

3.1. Análisis exploratorio de los datos.

Se generó una base de datos de precipitación con los datos diarios de estaciones meteorológicas pertenecientes al Servicio Meteorológico Nacional (SMN) que reportan en tiempo operacional a la red de la Organización Meteorológica Mundial (OMM) cubriendo la región Pampeana. Las mismas se seleccionaron teniendo en cuenta que tuvieran registros completos en el periodo de estudio 1979-2016, en este proceso previo de selección se eliminaron estaciones meteorológicas que tuvieran más del 10% de registros faltantes en los días del verano, para que no afecten el análisis de los forzantes climáticos y por lo tanto la construcción de los modelos de pronóstico que es el objetivo de este trabajo. En este proceso se eliminaron 5 estaciones meteorológicas en la región, que no estaban ubicadas en áreas críticas para este estudio, es decir, no afectaron el agrupamiento regional de las estaciones de medición, permitiendo que no queden áreas muy extensas sin cubrir sobre la región de estudio y teniendo en consideración que las que quedaran tengan una distribución geográfica relativamente uniforme en el área considerada. En general se puede decir que la distancia media entre estaciones en la región fue del orden de 100 km. Una vez conformada la base de datos diaria se calcularon los valores acumulados mensuales y estacionales de verano(DEF), otoño(MAM), invierno(JJA) y primavera(SON), trabajando luego solamente con la precipitación de verano.

La Tabla 1 detalla la lista de las 37 estaciones meteorológicas utilizadas, en particular se informa su identificación dentro de la Organización Meteorológica Mundial (idOMM), su nombre, latitud, longitud y elevación sobre el nivel del mar.

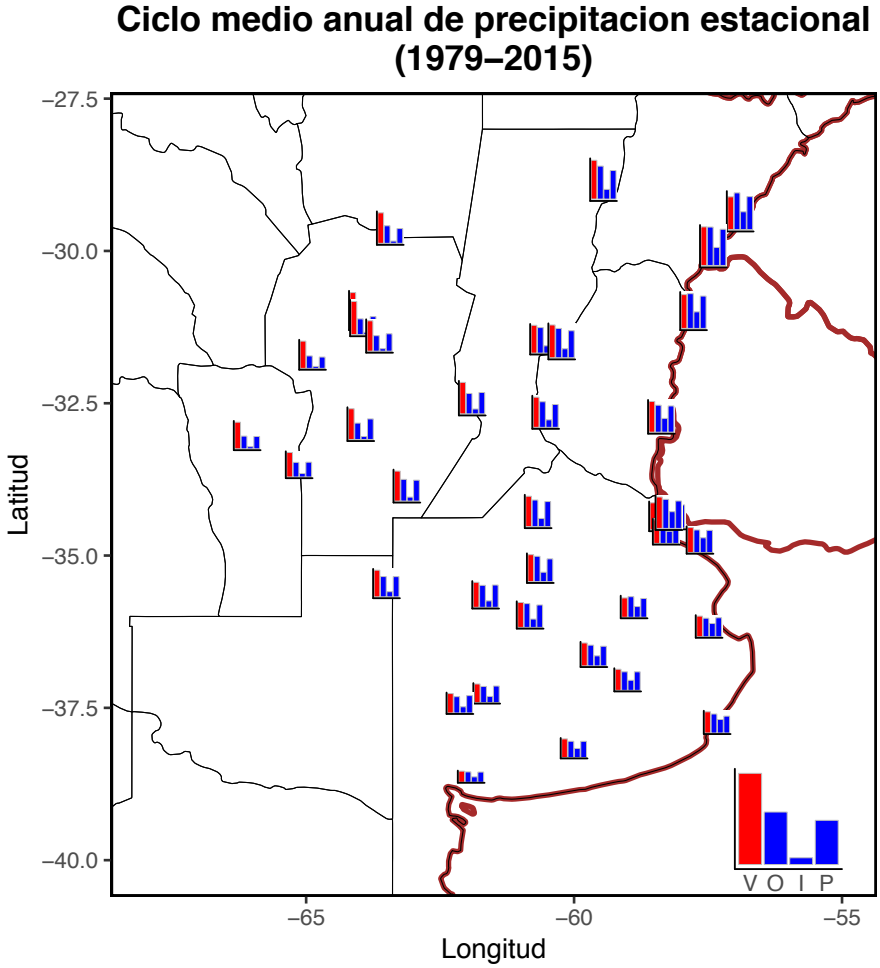


Figura 1. Localización de las estaciones meteorológicas y ciclos medios de precipitación estacionales en el periodo 1979-2015 (V-verano, O-otoño, I-invierno, P-primavera)

La Figura 1 muestra la climatología de la precipitación estacional de verano(V), otoño(O), invierno(I) y primavera(P) mapeado por estación meteorológica. En el mapa se muestra por localización de cada estación meteorológica y en esa posición un histograma donde cada barra representa la precipitación media de verano (rojo), otoño, invierno y primavera (azul) del período de años considerado (1979-2015). En el mapa se puede observar que hay un patrón de menores lluvias de verano al sur que se incrementan hacia el norte. Este patrón se puede observar en la zona sur de la provincia de Buenos Aires y la zona de la provincia de Corrientes como contrastantes. Por otro lado, se puede observar que las lluvias de invierno (letra I en la referencia del gráfico) van decreciendo de Este a Oeste como se observa en la región de CABA y la región de la provincia de San Luis como zonas contrastantes. Por lo tanto, estos comportamientos se observaron en las series representativas de los agrupamientos de las estaciones.

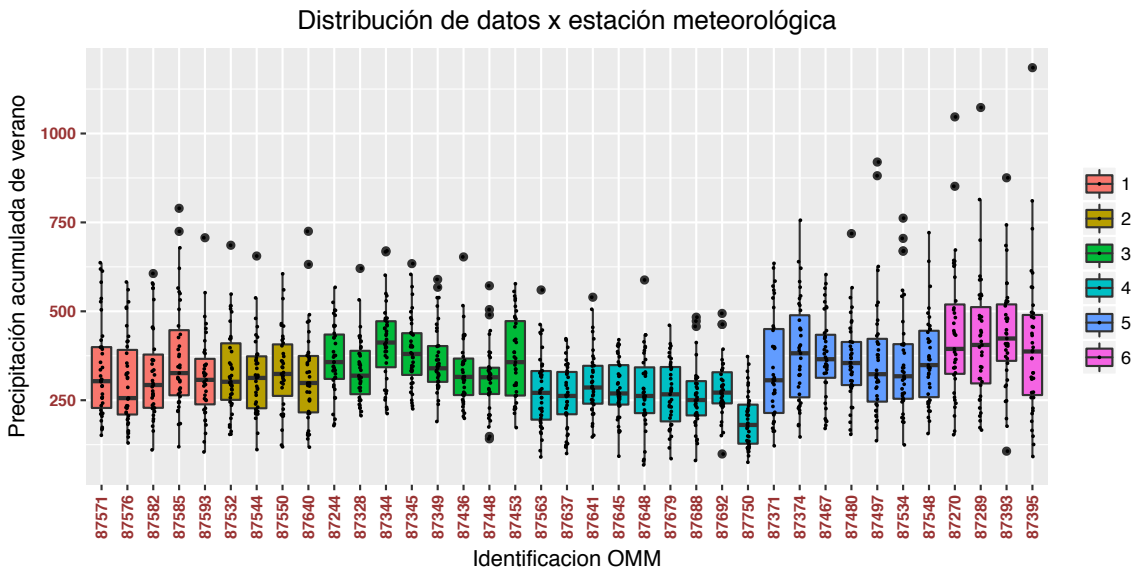


Figura 2. Boxplots de las estaciones meteorológicas utilizadas (colores indican agrupamiento de estaciones)

La Figura 2 muestra las distribuciones de las observaciones de precipitación acumulada de verano utilizando boxplots para cada una de las 37 estaciones meteorológicas consideradas, permitiéndonos observar los rangos de variabilidad de las mismas, para una mejor visualización de los datos por estación meteorológica se muestran los puntos dentro de cada boxplot. Las estaciones o boxplot se pintaron de color correspondientes a agrupamientos que se explicaran con posterioridad.

Se observaron algunos detalles interesantes en la Figura 2:

- Las estaciones meteorológicas pintadas de color turquesa corresponden a la región del Sur de la región considerada, que son las menores precipitaciones medias observadas en el periodo considerado.
- Las estaciones de color rosa corresponden a la región Norte de la región considerada, que son las de mayores precipitaciones medias observadas en el periodo considerado.
- En general se puede observar que existe una mayor cantidad de casos de precipitación más alta que la normal que de precipitación más baja.
- Los casos de precipitación más alta predominan en las estaciones meteorológicas de color rosa que están ubicadas al norte de la región Pampeana y son debidos a que la región es muy afectada por el fenómeno del ENSO “El Niño-Oscilación Sur” produciendo precipitaciones de verano en exceso en algunos años.
- Es importante destacar la alta variabilidad de la precipitación de verano en toda la región de estudio.

3.1.1. Agrupamientos de estaciones meteorológicas (Clusters).

Se aplicaron distintas técnicas de agrupamientos jerárquico y no jerárquico que se consideraron adecuadas para este problema, combinándolas del siguiente modo:

3.1.1.1. Agrupamiento Jerárquico:

Matrices de distancia:

- Euclídea
- Correlación

Métodos aglomerativos:

- Single linkage: Calcula todas las diferencias de pares entre los elementos en el grupo 1 y los elementos en el grupo 2, y considera la más pequeña de estas diferencias como un criterio de vinculación. Tiende a producir clusters alargados y sueltos.
- Complete linkage: Calcula todas las diferencias de pares entre los elementos del clúster 1 y los elementos del clúster 2, y considera el valor más grande (es decir, el valor máximo) de estas diferencias como la distancia entre los dos clústeres. Tiende a producir clusters más compactos.
- Average: Calcula todas las diferencias de pares entre los elementos en el grupo 1 y los elementos en el grupo 2, y considera el promedio de estas diferencias como la distancia entre los dos grupos.
- ward.D : Minimiza la varianza total dentro del clúster.

Se buscaron agrupamientos de 4 a 7 clases que permitieran discriminar aceptablemente bien los regímenes de precipitación de verano en la región Pampeana utilizando esta metodología jerárquica. El mejor resultado se obtuvo usando una matriz de distancia euclídea y el método aglomerativo de mínima varianza. Los gráficos de todas los experimentos se puden ver en el Anexo I.
La Figura 3 muestra el mejor resultado obtenido de acuerdo a los resultados esperados del agrupamiento.

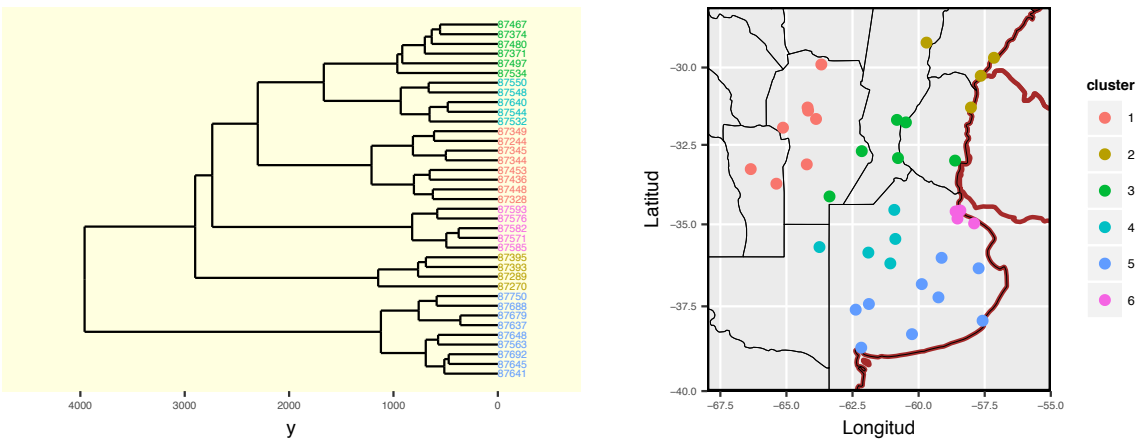


Figura 3. Mejor agrupamiento jerarquico usando distancia euclideana y método aglomerativo ward.D de minimización de la varianza dentro del grupo.

3.1.1.2. Agrupamiento No Jerárquico (K-Means):

Se utilizó la metodología de K-Means cuyo objetivo es encontrar agrupamientos en los datos, con el número de grupos previamente asignado por la variable K (cantidad de grupos). El algoritmo en general funciona iterativamente para asignar cada punto de

datos a uno de los grupos K en función de las características que se proporcionan. Los puntos de datos se agrupan según la similitud de características. Se utilizó la medida de validación interna SSW (Sum of Squared Within) para evaluar la cohesión de los grupos también llamada técnica de "Elbow" que considera la suma de los cuadrados interna del agrupamiento hasta que la misma deja de variar, punto en que no tiene sentido continuar agregando grupos. En este caso, el número de grupos se varió entre 1 y 10.

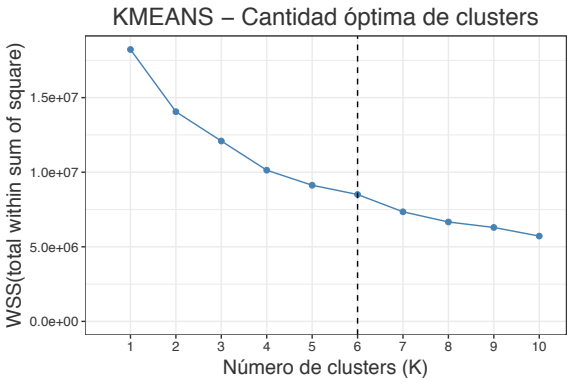


Figura 4. Metodo "Elbow" K optima

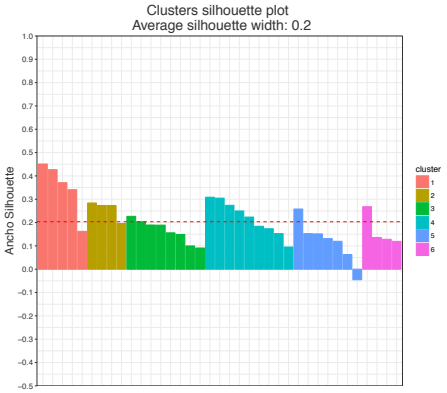


Figura 5. Silhouette del agrupamiento

Se consideró de acuerdo a lo que se muestra en la figura 4 que el número óptimo para el K podría estar entre 5 y 6 y se eligió un K de 6. La figura 5 muestra un gráfico de validación de cohesión y separabilidad denominado silhouette del agrupamiento para el caso de K igual a 6 donde se aprecia que los grupos están bastante bien conformados.

La Figura 6 muestra el resultado final del agrupamiento de 6 grupos, se puede observar que se verifica que los grupos definen zonas respetando aspectos de la precipitación conocida en cuanto a los gradientes de precipitación de sur a norte y de este a oeste, delimitando la zona central aproximadamente en la región de la zona denominada núcleo. En esta figura se puede observar el idOMM de identificación de las estaciones meteorológicas el que se puede asociar con la Tabla 1 que tiene los metadatos de las estaciones.

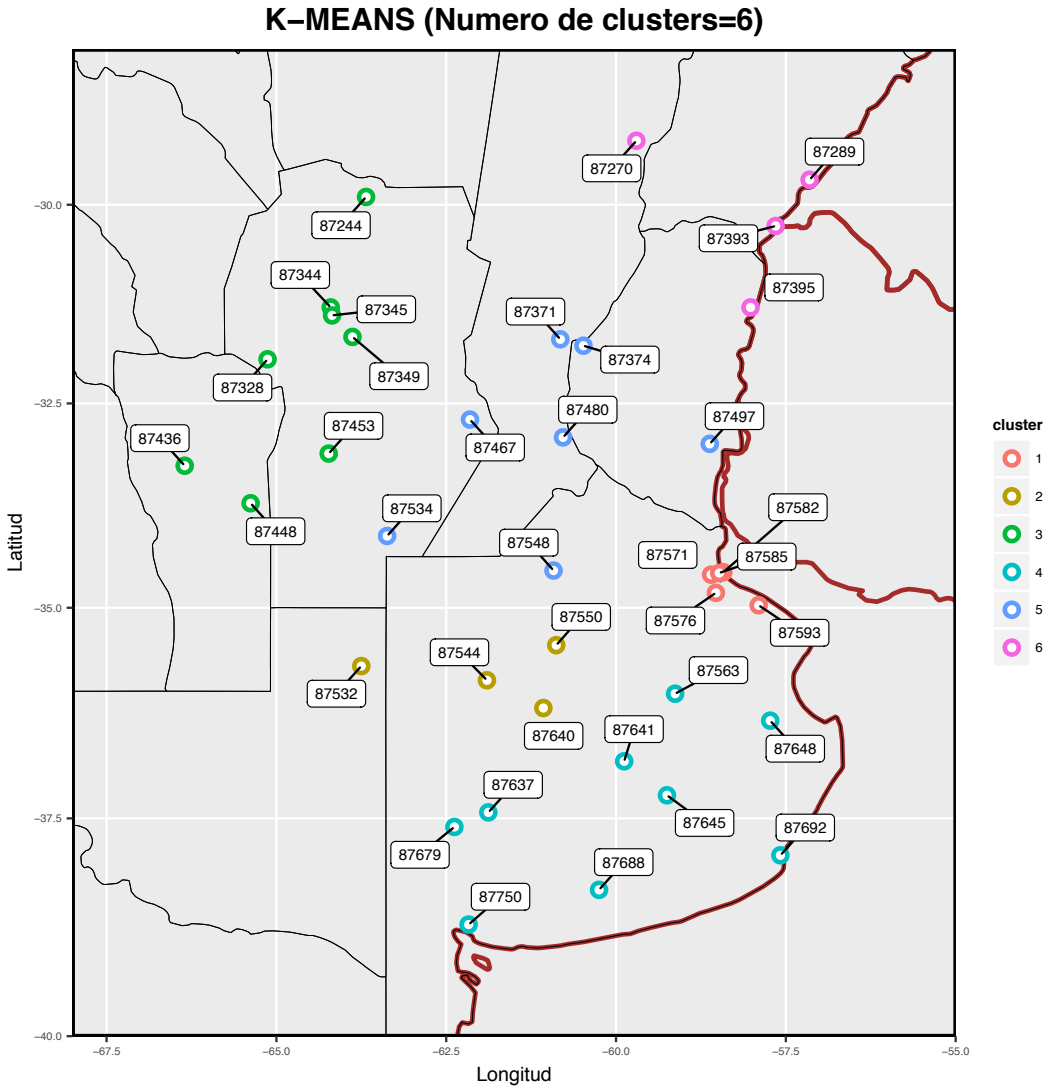


Figura 6. Agrupamiento usando K-Means (K = 6)

3.1.1.3. Análisis de Varianza de los agrupamientos (ANOVA).

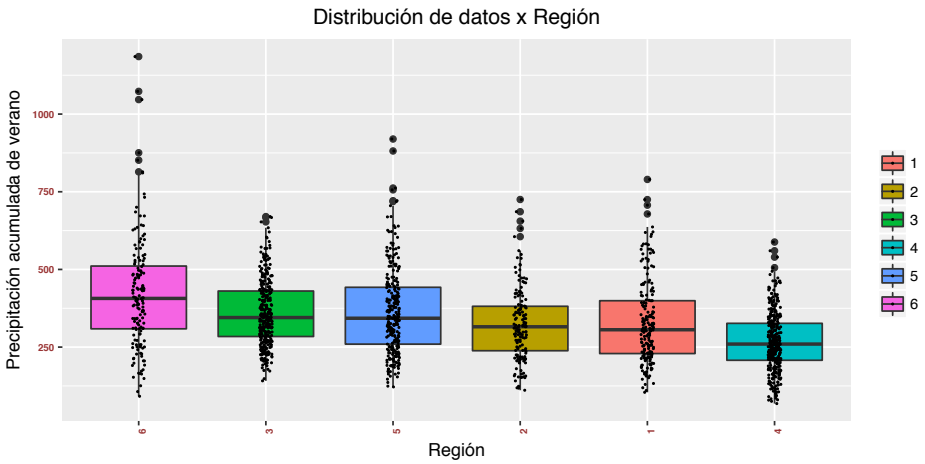


Figura 7. Boxplots de los agrupamiento por Región

Obtenida la regionalización, se realizó un análisis de varianza, para verificar si las diferencias entre los grupos encontrados son significativamente diferentes.
El resultado del test ANOVA fue:

```
                  Df   Sum Sq Mean Sq F value Pr(>F)
factor(cluster)    5  3324587  664917   42.14 <2e-16 ***
Residuals       1437 22675735   15780
```

El resultado fue el rechazo de la la hipótesis nula de igualdad de medias, (p-valor < 0.05), lo que indicaría que hay evidencia suficiente acerca que las medias de los grupos de las regiones no son iguales.
En vista de este resultados, para verificar si el resultado es válido se verificaron los supuestos del método (independencia, normalidad, homocedasticidad). La normalidad de las distribuciones se probó utilizando el test de Shapiro.

```
        Shapiro-Wilk normality test

data:  residuals(res.aov)
W = 0.96197, p-value < 2.2e-16
```

El resultado del test indica que las muestras no son normales (p-valor < 0.05).
Entonces se aplicó una transformación de Box-Cox para cumplir con el supuesto de normalidad. La transformación a aplicar es 0.2626263.

```
        Shapiro-Wilk normality test

data:  residuals(res.aov)
W = 0.99897, p-value = 0.5922
```

Una vez corregida la normalidad de las muestras, se aplicó el test de Levene para testear homogeneidad de varianzas (Homocedasticidad).

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    5  6.7119 3.422e-06 ***
      1437
```

Se observó que no se cumple con homogeneidad de varianzas y por lo tanto el test paramétrico de ANOVA aplicado no cumple con los supuestos, por lo tanto, se lo consideró no valido.
Como método alternativo, se aplicó un test de ANOVA no parametrico para muestras independientes denominado de Krukal-Wallis (Wilcoxon).

```
        Kruskal-Wallis rank sum test

data:  DataEst_v$DEF^(trf) and factor(DataEst_v$cluster)
Kruskal-Wallis chi-squared = 178.01, df = 5, p-value < 2.2e-16
```

Se verificó con el metodo no paramétrico que hay diferencias significativas entre grupos y también se aplicó el mismo test para hacer multicomparaciones entre grupos con un p.valor de 0.05.

```
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
      obs.dif critical.dif difference
1-2   3.53141    131.38256      FALSE
1-3 137.89519    111.65370       TRUE
```

```
1-4 190.48462    109.24189       TRUE
1-5 116.66923    114.68010       TRUE
1-6 254.99615    131.38256       TRUE
2-3 134.36378    119.93532       TRUE
2-4 194.01603    117.69334       TRUE
2-5 113.13782    122.75771      FALSE
2-6 251.46474    138.48938       TRUE
3-4 328.37981     95.16772       TRUE
3-5  21.22596    101.36384      FALSE
3-6 117.10096    119.93532      FALSE
4-5 307.15385     98.70091       TRUE
4-6 445.48077    117.69334       TRUE
5-6 138.32692    122.75771       TRUE
```

Los resultados evidencian que "en general" los grupos son significativamente diferentes utilizando esta metodología no paramétrica, y que los grupos detectados por este método si bien no tiene en cuenta la distancia o posición geográfica entre grupos, detectaron diferencias entre grupos alejados entre sí y como similares algunos grupos que son vecinos geográficamente o mejor dicho, que tienen límites vecinos, aunque desde el punto de vista de influencia de los forzantes climáticos estos grupos podrían ser regiones distintas.

3.1.1.4. Armado de series medias regionales de precipitación.

Una vez, definida la regionalización para el área Pampeana, se calcularon las series medias estacionales por área, las mismas se muestran en la Tabla 2 y los agrupamientos a los que corresponden se pueden observar en la Figura 6. Estas series son las que se correlacionaron con las 7 (siete) variables meteorológicas del mes de noviembre: altura geopotencial en diferentes niveles de la atmósfera, agua disponible en la columna de agua, viento en capas bajas y temperatura de superficie del mar globales que se transformaran en predictores para los modelos de regresión lineal múltiple regionales.

3.1.2. Análisis de forzantes globales.

Como se detalló anteriormente las variables que se usaron como forzantes globales fueron la altura geopotencial (HGT) en 1000, 500 y 200 hPa el viento zonal (U) y meridional (V) en 850 hPa, agua precipitable (TCW) en la capa desde superficie hasta 700 hPa y la temperatura de superficie del mar (SST) obtenidas de los reanálisis de NCEP/NCAR, estos reanálisis son información de tipo matricial y tienen una resolución espacial aproximada de 250 Km (2.5 grados) y una resolución temporal que en este caso es mensual. En particular se utilizó el mes previo al verano (noviembre) para realizar la correlación con las precipitaciones de verano. La dimensión de estas matrices es de 144 puntos de longitud x 73 puntos de latitud x 37 tiempos (noviembres) por variable.

La metodología aplicada calcula la correlación lineal de las series mensuales de los forzantes en cada punto de retículo en el periodo (1979-2015) del mes de noviembre contra las series de observaciones medias de cada uno de los agrupamientos de la región Pampeana. El resultado fueron 7 mapas de correlación uno por cada variable /forzante por cada agrupamiento de la región pampeana, es decir, 42 mapas de correlación. En estos mapas la correlación significativa al 95% en el umbral mayor es igual a 0.33 debido a que en este caso el número de elementos de la serie N es de 37 años.

Normalmente para la tarea de delimitar las regiones que son significativamente interesantes por su correlación y de donde se extraen las series de predictores, se asignan a "cajas" o retículos por rangos de latitud y longitud. Para este trabajo se

decidió apartarse de esta metodología e implementar un algoritmo para delimitar el convexo de correlación significativa guardando los puntos que lo componen y armando la serie temporal areal del predictor usando solo los puntos de la matriz que superan el umbral de correlación significativa.

La Figura 8 a modo de ejemplo muestra los siete (7) mapas de correlaciones de la región 5 (cinco), donde la escala es roja para correlaciones positivas y azules para las negativas. El resto de los mapas de correlaciones no se muestran en este informe, pero se puede acceder a ellos en el GitHub del trabajo:

(https://github.com/alrolla/Especializacion_2018/tree/master/Predictores).

Describiremos sintéticamente las regiones globales donde los forzantes están asociados con la variabilidad de la precipitación en el Sudeste de Sudamérica (SESA).

Los forzantes en general que se consideraron para este trabajo se definieron en el hemisferio Sur, los mismos se pueden describir del siguiente tipo:

- Relacionadas con la temperatura de la superficie del mar (SST): Las regiones interesantes en este caso son las del Pacífico ecuatorial (ENSO: El Niño - Southern Oscillation)), Océano Indico (IND), Atlántico Sur (ATL_S) y Pacífico Sur (PAC_S).

- Dinámicos: Correspondiente a variables de altura geopotencial (HGT100, HGT500, HG200). Las regiones interesantes en este caso son las del Pacifico Sur (PAC_S), Indico (IND), Atlántico Sur (ATL_S) y viento en capas bajas (U850) Viento en la región de Pacifico (PAC_S), Indico (IND), Atlántico sur (ATL_S).

- Anticiclón del Atlántico: Correspondiente a las variables de altura geopotencial y viento en la zona del anticiclón permanente del Atlántico Sur, ubicado en la región subtropical.

- Agua precipitable: Correspondiente al total de agua precipitable disponible en la columna sobre la región Pampeana.

- LLJ (Low Level Jet): Correspondiente a la variable de viento meridional (V850) sobre el Norte Argentino.

Se describirá en detalle el procesamiento que se realizó para una de las seis (6) regiones definidas el resto de los gráficos y resultados están disponibles en el GitHub del trabajo.

Las tablas 3 a la 8 muestran los predictores (forzantes) seleccionados para cada una de las 6 regiones de la zona Pampeana en base a los mapas de correlación.

Tabla 3. Predictores seleccionados de la región 1

Variables	Predictores			
SST				
HGT1000	HGT1000-10	HGT1000-13		
HGT500	HGT500-4			
HGT200	HGT200-3			
TCW				
U850	U850-5			
V850	V850-9			

Tabla 4. Predictores seleccionados de la región 2

Variables	Predictores			
SST	SST-5			
HGT1000	HGT1000-2	HGT1000-9	HGT1000-14	
HGT500				
HGT200	HGT200-2			
TCW				
U850	U850-10			
V850				

Tabla 5. Predictores seleccionados de la región 3

Variables	Predictores			
SST	SST-5	SST-21	SST-38	SST-46
HGT1000	HGT1000-2	HGT1000-5	HGT1000-6	
HGT500	HGT500-2	HGT500-3	HGT500-17	
HGT200	HGT200-2	HGT200-3	HGT200-4	HGT200-9
TCW				
U850	U850-6	U850-11	U850-12	U850-18
V850				

Tabla 6. Predictores seleccionados de la región 4

Variables	Predictores			
SST	SST-1			
HGT1000	HGT1000-1	HGT1000-11		
HGT500	HGT500-3			
HGT200	HGT200-5			
TCW				
U850	U850-15	U850-16		
V850				

Tabla 7. Predictores seleccionados de la región 5

Variables	Predictores			
SST	SST-8	SST-21	SST-25	SST-41
HGT1000	HGT1000-13	HGT1000-14	HGT1000-17	
HGT500	HGT500-4	HGT500-6	HGT500-8	
HGT200	HGT200-6	HGT200-9		
TCW	TCW-26			
U850	U850-10	U850-17	U850-23	
V850	V850-14			

Tabla 8. Predictores seleccionados de la región 6

Variables	Predictores			
SST	SST-8	SST-40	SST-65	SST-68
HGT1000	HGT1000-2	HGT1000-19		
HGT500	HGT500-2	HGT500-7		
HGT200	HGT200-4	HGT200-7	HGT200-9	HGT200-13
TCW	TCW-34			
U850	U850-4	U850-15	U850-17	U850-26
V850	V850-16			

3.2. Pre-selección de predictores(LASSO).

Para evitar los efectos adversos del problema de colinealidad de los predictores en un modelo lineal estimado por mínimos cuadrados, en el contexto $p < n$ se utilizó la técnica de regresión LASSO, que es muy similar a los mínimos cuadrados (OLS), a excepción de que los coeficientes se estiman minimizando una cantidad diferente. Motivado por el objetivo de encontrar una técnica de regresión lineal que, mediante la contracción de los coeficientes, lograra estabilizar las estimaciones y predicciones y que realizase SELECCIÓN de predictores, se aplicó la técnica LASSO (least absolute shrinkage and

SELECTION operator). Es una técnica de regresión lineal regularizada, similar a Ridge, con una leve diferencia en la penalización que trae consecuencias importantes.

En especial, a partir de cierto valor del parámetro de complejidad el estimador de Lasso produce estimaciones nulas para algunos coeficientes y no nulas para otros, con lo cual Lasso realiza una selección de variables en forma continua, debido a la norma L1 (valor absoluto). Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes y al mismo tiempo produce modelos interpretables por la reducción de algunos coeficientes a cero. En general, los modelos generalizados Lasso son mucho más fáciles de interpretar que los obtenidos mediante Ridge. Se buscó el mejor valor del parámetro λ por el método de validación cruzada. Algo muy interesante de Lasso es que si hay un grupo de variables entre las cuales las correlaciones por parejas son muy altas, entonces Lasso tiende a seleccionar solo una variable del grupo, sin importarle cuál de ellas selecciona.

Para aplicar el método de LASSO se estandarizaron previamente las variables. La figura 7 muestra el gráfico resumen de aplicar LASSO haciendo "cross-validation" del tipo "Leave-one-out cross-validation" (LOOCV) utilizando los datos de la región 5. Se detalla la curva de validación cruzada (MSE: mean square error) como una línea punteada roja y las curvas de desviación estándar superior e inferior a lo largo de la secuencia de λ (barras de error). Las líneas punteadas verticales indican la banda (punteadas el λ mínimo y el primer λ que supera en un desvío al λ mínimo) en la que los predictores tienen un MSE bajo, es decir, indica la cantidad y los predictores significativos para la regresión. En el eje superior se ve la cantidad de coeficientes distintos de cero en ese intervalo.

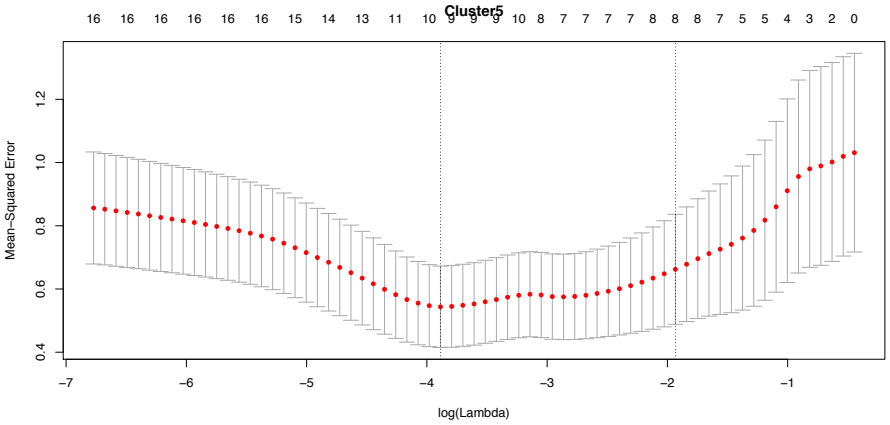


Figura 7. Región 5. "Leave-one-out" cross-validation.

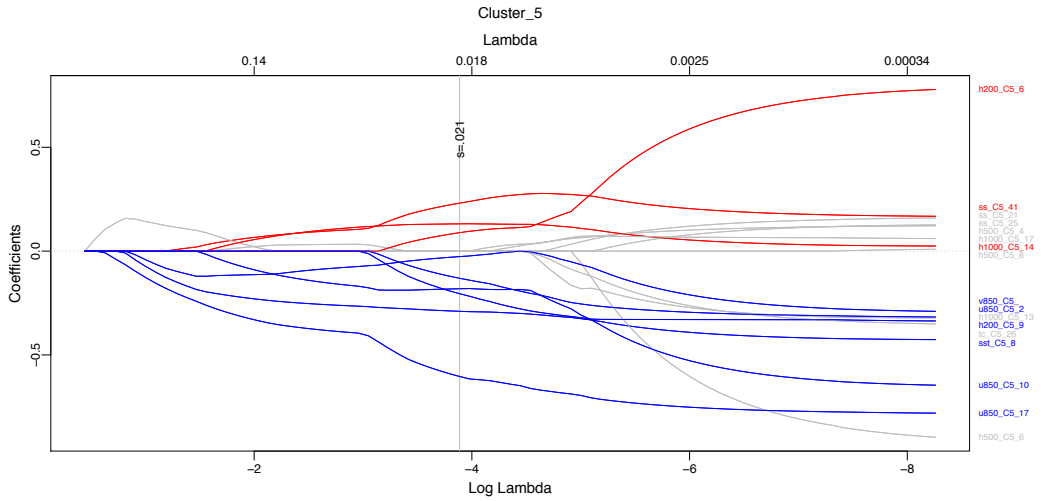


Figura 8. Región 5. Gráfico de coeficientes y predictores seleccionados por LASSO.

La figura 8 muestra los predictores seleccionados para λ cuyo MSE es mínimo, en el caso mostrado es de (λ=0.21), correspondiente a los predictores de colores rojo y azul, los pintados en gris se los consideró predictores no significativos, es decir, no aportan información adicional para la regresión.

Este método de selección de predictores se aplicó a todos los agrupamientos. Los gráficos de Lasso asociados a las restantes regiones se pueden ver en el GitHub asociado (**https://github.com/alrolla/Especializacion_2018/tree/master/Lasso**). Los predictores seleccionados se ven reflejados en las tablas de la 3 a la 8, los predictores (celdas de la tabla) griseados son los predictores que el método descartó para construir las regresiones lineal múltiples.

3.3. Construcción de los modelos.

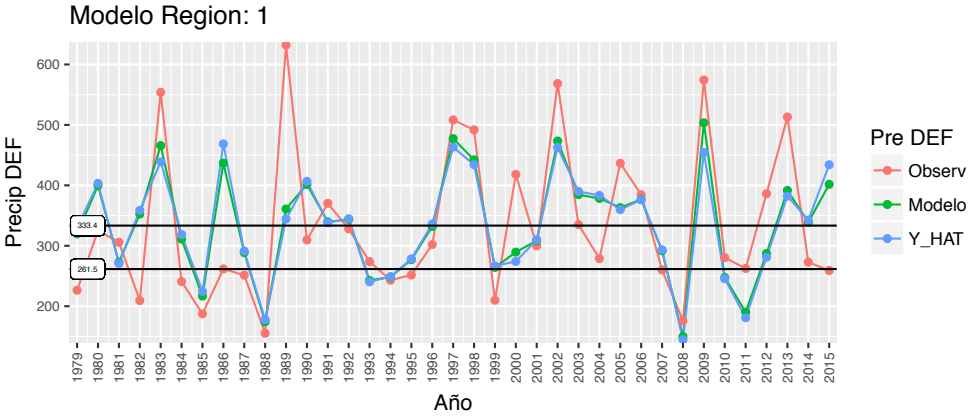
Debido a que para el correcto funcionamiento del método de LASSO se deben normalizar las variables, los coeficientes de la regresión pierden las unidades originales de medida de los predictores, como en el método de regresión múltiple estándar.

Por lo tanto, luego de haber obtenido los predictores que "describen mejor el problema" en cada región usando el método de LASSO, se aplicó un método automático de construcción de los modelos en base a esos predictores usando las unidades originales de los predictores. Se eligió el método "stepwise regression" que, en cada paso, considera una variable que se la suma o la resta del conjunto de variables explicativas en función de un criterio pre-especificado. De esta forma se generaron varios modelos cada uno de los cuales se deriva de la aplicación del método a diferentes conjuntos de predictores independientes.

Una vez obtenido el conjunto de todos los modelos posibles, se utilizó como criterio de selección del mejor modelo a la métrica de maximizar AdjR2 (R2 ajustado), lo cual implica maximizar la varianza de la precipitación estival explicada por el predictor. Todos los modelos con las métricas asociadas de CV y adjR2 se pueden ver en GitHub (**https://github.com/alrolla/Especializacion_2018/tree/master/Modelos**).

3.4. Verificación de los modelos.

A continuación, se muestran los análisis de cada uno de los modelos para cada región, junto a su verificación. En todos los casos la figura superior muestra las series de observaciones (línea roja), la serie derivada de la aplicación del mejor modelo (línea verde) y la serie derivada de la evaluación respecto de la validación cruzada "Leave-one-out cross-validation" (LOOCV) (línea azul, Y_HAT). Al pie de la figura se muestra la ecuación del modelo y la varianza explicada por el mismo (adjR2). Se observan además dos líneas horizontales (líneas negras) que representan los umbrales de los terciles de las observaciones. Usando estos terciles se clasificaron las observaciones en Sobre-normal, Normal y Sub-normal, definiendo así las clases y se confeccionó una tabla de contingencia con estas 3 clases. Además, se confeccionaron 3 tablas de confusión una para cada clase y se calcularon los índices de **recall** (probability of detection), **FNR** (false alarm) y **accuracy** (sensitivity).



R1_PPverano = *-712,58 + 10,71 * **hgt1000_C1_13** - 15,63 * **u850_C1_5** - 21,07 * **v850_C1_9***
(AdjR2 = 0,447)

R1		OBS		
		sub	normal	sobre
MOD	sub	4	3	0
	normal	6	2	2
	sobre	2	7	11

SUBNORMAL		
	YES	NO
YES	4	3
NO	8	22

ACCURACY	RECALL	FNR
0,7	0,33	0,12

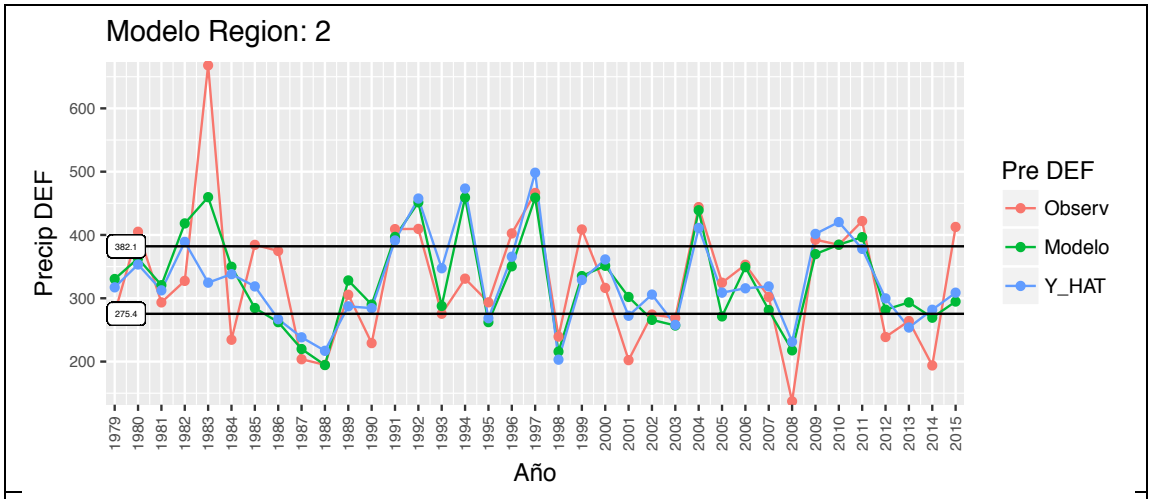
NORMAL		
	YES	NO
YES	2	8
NO	10	17

ACCURACY	RECALL	FNR
0,51	0,17	0,32

SOBRENORMAL		
	YES	NO
YES	11	9
NO	2	15

ACCURACY	RECALL	FNR
0,7	0,85	0,38

Región 1: Se ve que el modelo construído tiene problemas para representar los años extremos de precipitación de verano, sólo representa el 45% de la variabilidad de las observaciones. Los predictores/forzantes climáticos que quedaron en el modelo como forzantes de la precipitación fueron la altura geopotencial en capas bajas en el Indico (hgt1000), los vientos del oeste (u850) en el Pacifico sur y el viento meridional en la zona del LLJ en el continente sudamericano entrantes a la región. Los extremos de Sobre-Normal de precipitación corresponden con "años Niño", como se observa la temperatura de superficie del mar no fue seleccionado por el método como predictor dado que no mostró una región con correlación significativa con la precipitación de la región 1. Con respecto a la capacidad de clasificación del modelo se observa que para Sub-normal y Sobre-normal tiene un **accuracy** bueno y para la clase Normal decrece. El mejor desempeño lo tiene para la clase sobre-normal con un **accuracy** de 0.7 y un **recall** de 0.85, aunque con un índice no muy bueno de falsa alarma que asciende a 0.38 (**FNR**).



R2_PPverano** = 19626,02 - 70,34 * **sst_C2_5** + 1,02 * **hgt1000_C2_2** + 1,22 * **hgt1000_C2_9** - 1,39 * **hgt200_C2_2
(AdjR2 = 0,477)

R2		OBS		
		sub	normal	sobre
MOD	sub	7	2	0
	normal	5	8	7
	sobre	0	2	6

SUBNORMAL		
	YES	NO
YES	7	7
NO	5	23

ACCURACY	RECALL	FNR
0,81	0,58	0,08

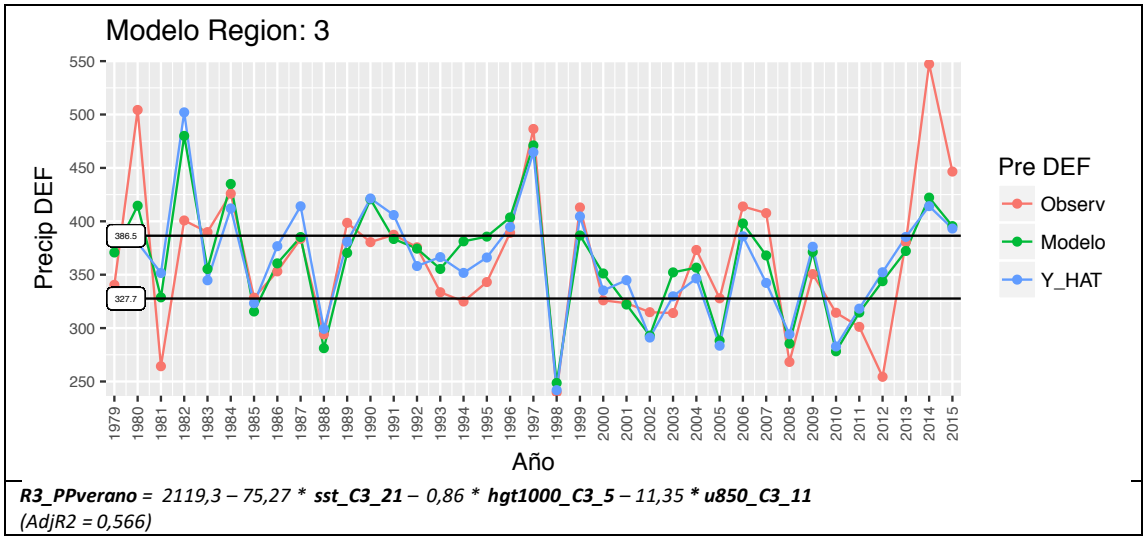
NORMAL		
	YES	NO
YES	8	12
NO	4	13

ACCURACY	RECALL	FNR
0,57	0,67	0,48

SOBRENORMAL		
	YES	NO
YES	6	2
NO	7	22

ACCURACY	RECALL	FNR
0,76	0,46	0,08

Región 2: El modelo es capaz de representar sólo el 48% de la variabilidad de las observaciones. Los predictores/forzantes climáticos que conforman el mejor modelo fueron la temperatura de la superficie del mar en el Pacífico occidental y la altura geopotencial en capas bajas en el Pacifico y Atlántico sur y en capas altas en el Indico. El extremo del año 1983 corresponde a un año *Niño* muy fuerte, en general los extremos de precipitación en la clase Sobre-Normal corresponden a años *Niño* (aunque no se corresponden con la intensidad de los mismos). Con respecto a la capacidad de clasificación del modelo, podemos decir, que para Sub-normal y Sobre-normal tiene un **accuracy** interesante y para la clase Normal es más bajo. El **recall** de cada clase nos indica que el modelo detecta más del 50% de los casos de los posibles correspondientes a cada una de las clases. La tasa más alta de clasificación errónea se produce en la clase Normal (FNR) y asciende al 48%.



R3_PPverano = *2119,3 – 75,27 * **sst_C3_21** – 0,86 * **hgt1000_C3_5** – 11,35 * **u850_C3_11***
(AdjR2 = 0,566)

R3		OBS		
		sub	normal	sobre
MOD	sub	6	2	0
	normal	6	7	5
	sobre	0	2	8

SUBNORMAL		
	YES	NO
YES	6	2
NO	6	23

ACCURACY	RECALL	FNR
0,78	0,50	0,08

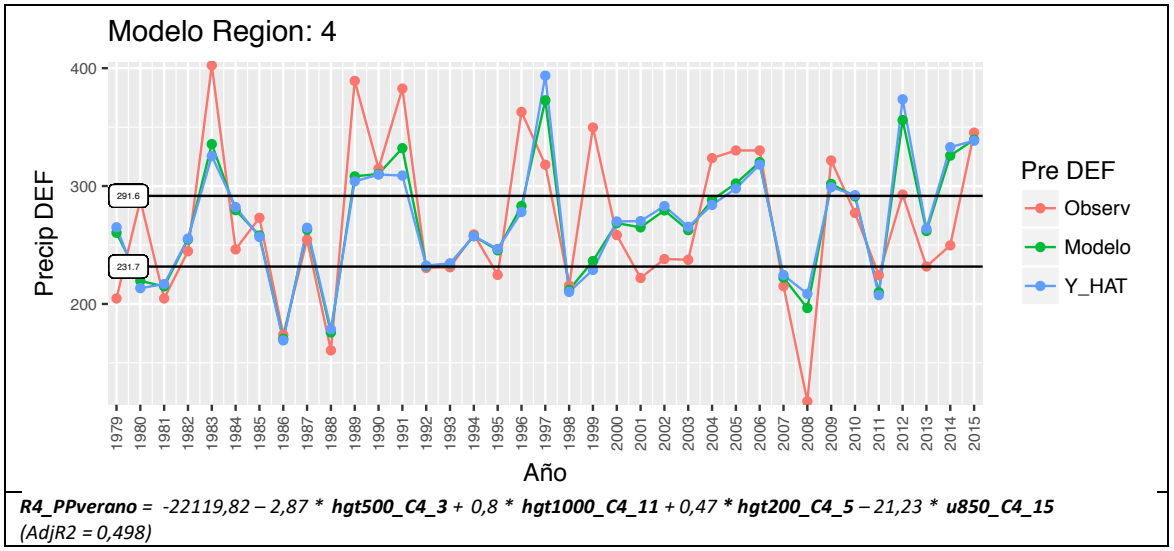
NORMAL		
	YES	NO
YES	7	11
NO	5	14

ACCURACY	RECALL	FNR
0,57	0,58	0,44

SOBRENORMAL		
	YES	NO
YES	8	3
NO	5	21

ACCURACY	RECALL	FNR
0,78	0,62	0,12

Región 3: El modelo es capaz de representar el 57% de la variabilidad de las observaciones. Los predictores/forzantes climáticos que construyeron el mejor modelo fueron los correspondientes a la temperatura de la superficie del mar en el Pacífico central, la altura geopotencial en capas bajas en el Índico y el Viento zonal en la Atlántico sur. La región 3 es una zona particular ya que posee importante orografía. Con respecto a la capacidad de clasificación del modelo, podemos decir, que para Sub-normal y Sobre-normal tiene un **accuracy** alto, aunque para la clase Normal es menor. El **recall** de cada clase nos indica que el modelo clasifica bien más del 50% de los casos correspondientes a cada una de ellas. La tasa más alta de clasificación errónea se produce en la clase Normal (FNR) y asciende al 44%.



R4		OBS		
		sub	normal	sobre
MOD	sub	7	1	1
	normal	5	9	2
	sobre	0	2	10

SUBNORMAL		
	YES	NO
YES	7	7
NO	5	23

ACCURACY	RECALL	FNR
0,81	0,58	0,08

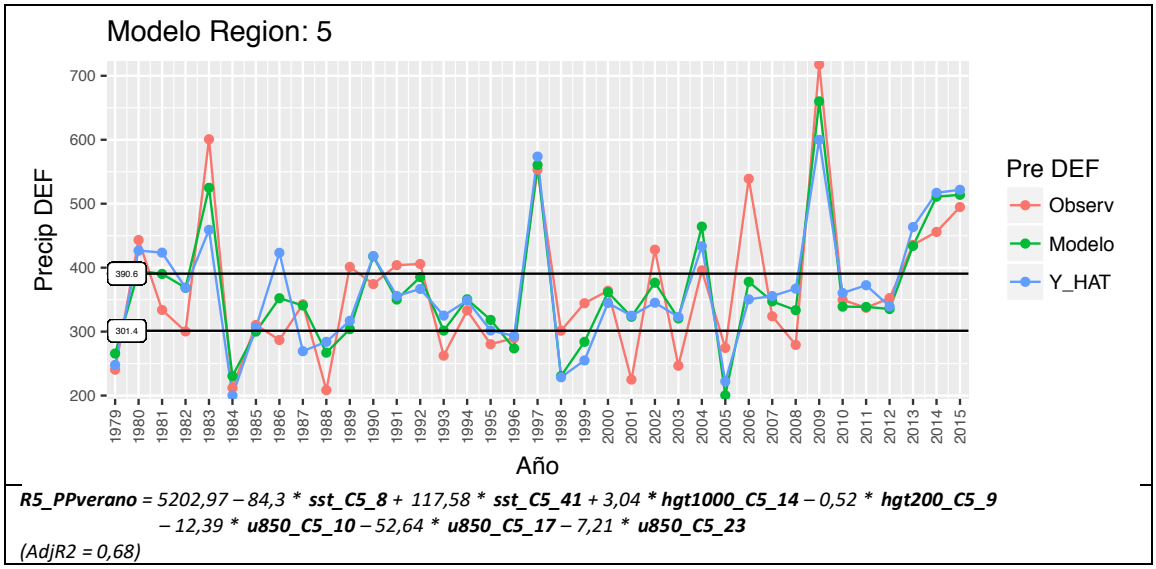
NORMAL		
	YES	NO
YES	8	12
NO	4	13

ACCURACY	RECALL	FNR
0,73	0,75	0,28

SOBRENORMAL		
	YES	NO
YES	6	2
NO	7	22

ACCURACY	RECALL	FNR
0,86	0,77	0,08

Región 4: El modelo es capaz de representar el 50% de la variabilidad de las observaciones. Los predictores/forzantes climáticos que construyeron el modelo fueron la altura geopotencial en capas medias en el Índico, en capas bajas en el Atlántico sur y en capas altas en el Pacífico central y el Viento zonal en la zona de entrada del anticiclón del Atlántico sobre el continente en Sudamérica. La región 4, corresponde al sur de la provincia de Buenos Aires, es la zona donde se registran las menores precipitaciones. Con respecto a la capacidad de clasificación del modelo, se observa que para las tres clases tiene un **accuracy** bueno, aunque para la clase Normal es menor. El **recall** de cada clase nos indica que el modelo clasifica bien más del 50% de los casos de cada una de ellas. La tasa más alta de clasificación errónea se registra en la clase Normal (FNR) y es del 28%.



R5_PPverano *= 5202,97 – 84,3 * **sst_C5_8** + 117,58 * **sst_C5_41** + 3,04 * **hgt1000_C5_14** – 0,52 * **hgt200_C5_9** – 12,39 * **u850_C5_10** – 52,64 * **u850_C5_17** – 7,21 * **u850_C5_23***

(AdjR2 = 0,68)

R5		OBS		
		sub	normal	sobre
MOD	sub	5	3	0
	normal	6	7	5
	sobre	1	2	8

SUBNORMAL		
	YES	NO
YES	5	3
NO	7	22

ACCURACY	RECALL	FNR
0,73	0,42	0,12

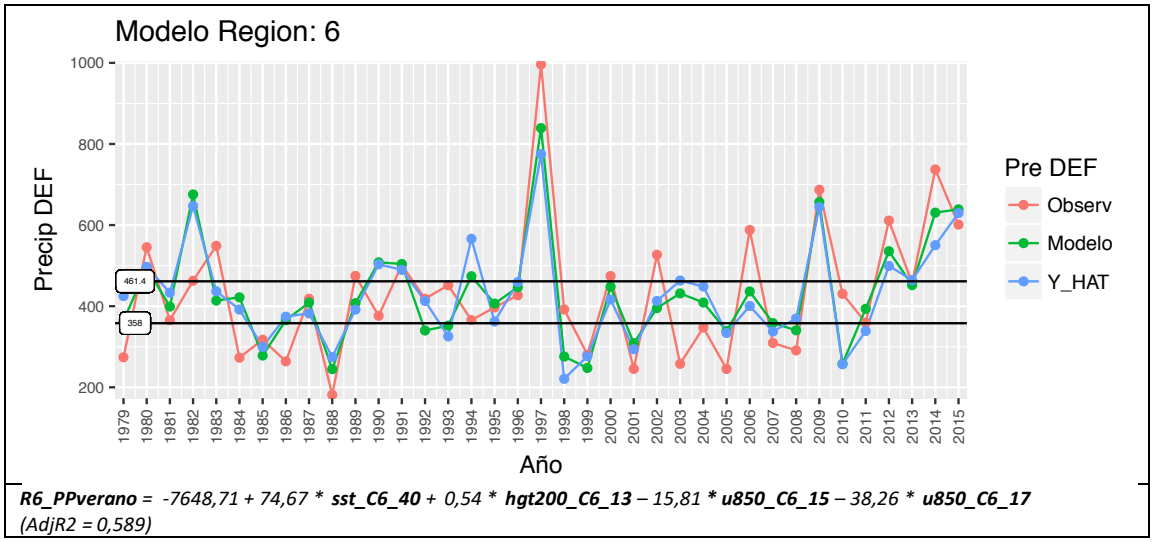
NORMAL		
	YES	NO
YES	7	11
NO	5	14

ACCURACY	RECALL	FNR
0,57	0,58	0,44

SOBRENORMAL		
	YES	NO
YES	8	3
NO	5	21

ACCURACY	RECALL	FNR
0,78	0,62	0,12

Región 5: El modelo es capaz de representar el 70% de la variabilidad de las observaciones. Los predictores/forzantes climáticos que construyeron el mejor modelo fueron la temperatura del mar en la región del Índico, la altura geopotencial en capas bajas en el Atlántico en la entrada del anticiclón, y en capas altas en el Atlántico sur y el Viento en la franja de los oestes del Pacífico y Atlántico sur. La región 5, corresponde a la región denominada zona "núcleo". Se puede observar en la figura que este modelo, como lo indica su R2, detecta muy bien los extremos de la serie. Con respecto a la capacidad de clasificación del modelo, se observa, que las tres clases tienen un **accuracy** aceptable, aunque para la clase Normal es menor. El **recall** de cada clase nos indica que el modelo clasifica bien los casos Sobre-normal, Sub-normal, Normal en orden decreciente del 40% a más del 60% de los casos en cada una de ellas, en particular la mejor es Sobre-normal. La tasa más alta de clasificación errónea se produce en la clase Normal (FNR) y asciende al 58%.



R6_PPverano = *-7648,71 + 74,67 * **sst_C6_40** + 0,54 * **hgt200_C6_13** – 15,81 * **u850_C6_15** – 38,26 * **u850_C6_17***
(AdjR2 = 0,589)

R6		OBS		
		sub	normal	sobre
MOD	sub	6	4	0
	normal	5	5	5
	sobre	1	3	8

SUBNORMAL		
	YES	NO
YES	6	4
NO	6	21

ACCURACY	RECALL	FNR
0,73	0,50	0,16

NORMAL		
	YES	NO
YES	5	10
NO	7	15

ACCURACY	RECALL	FNR
0,54	0,42	0,40

SOBRENORMAL		
	YES	NO
YES	8	4
NO	5	20

ACCURACY	RECALL	FNR
0,76	0,62	0,17

Región 6: El modelo es capaz de representar el 60 % de la variabilidad de las observaciones. Los predictores/forzantes climáticos que construyeron el mejor modelo fueron la temperatura del mar en la región del Pacífico ecuatorial y Pacífico sur, la altura geopotencial en capas altas en el Pacífico Occidentaly el Viento zonal en la franja sur del Indico y Atlántico. La región 6, corresponde a la región norte de la región Pampeana, es la zona donde se registran las mayores precipitaciones. Se puede observar en la figura que este modelo detecta bastante bien los extremos de la serie. Con respecto a la capacidad de clasificación del modelo, se observa que para las tres clases tiene un **accuracy** bueno, aunque para la clase Normal es menor. El **recall** de cada clase indica que el modelo clasifica bien Sobre-normal, Sub-normal, Normal en orden decreciente del 40% a más del 60% de los casos en cada una de ellas, en particular los mejores resultados se obtienen para la clase Sobre-normal. La tasa más alta de clasificación errónea se produce en la clase Normal (FNR) y es el 42%.

3.5. Pronóstico estacional de verano.

Para realizar la verificación de los modelos se utilizó el año 2016. La estrategia fue aplicar los seis modelos regionales construidos para la serie representativa de cada agrupamiento en base al período 1979-2015, a las series de precipitación de verano en las estaciones meteorológicas que componen cada grupo y calcular el valor correspondiente al año 2016. El resultado de la evaluación es un valor numérico de precipitación de verano que se clasificó usando los terciles de la serie histórica. Con esos umbrales de primer y segundo tercil, se clasificó en clases denominadas Sub-normal, Normal y Sobre-normal. Para ser objetivos con la clasificación a su vez se definió un índice de ajuste para la clasificación que establece que si la relación de

diferencia entre la clase observada y pronosticada es menor o igual a un solo salto de clase entre la clasificación observada y la pronosticada la clasificación se la considera **aceptable**. Si el salto fuera de dos clases la clasificación se la considera **no aceptable**. La tabla 9 muestra la combinación de Observación/Pronóstico de las clases y el código de colores utilizado:

Observado	Pronosticado	Verificación
Normal	Normal	Aceptable
Normal	Sobre-Nomal	Aceptable
Normal	Sub-Normal	Aceptable
Sub-Nomal	Sub-Nomal	Aceptable
Sub-Normal	Normal	Aceptable
Sub-Normal	Sobre-Nomal	No aceptable
Sobre-Normal	Sobre-Normal	Aceptable
Sobre-Normal	Normal	Aceptable
Sobre-Normal	Sub_normal	No aceptable

Tabla 9. Verificación de combinaciones de Observación/Pronóstico.

Las siguientes tablas muestran los resultados de la aplicación de los modelos a cada una de las estaciones que componen cada agrupamiento. Las columnas indican la identificación de la estación meteorológica, el valor observado de precipitación para el verano de 2016, el valor pronosticado por el modelo de la región 1, los umbrales inferior y superior de los terciles de las series observadas, la clase de la observación, la clase del pronóstico, el salto en la clasificación observación/pronostico y el acierto del modelo.

idOMM	Observado	Pronóstico	Tercil Inferior	Tercil Superior	Clase.OBS	Clase.PRON	Diferencia	Acierto
87571	342,0	471,9	276,1	328,6	sobre	sobre	0	SI
87576	354,3	388,3	234,2	318,6	sobre	sobre	0	SI
87582	362,4	431,0	267,2	327,4	sobre	sobre	0	SI
87585	344,0	492,3	307,5	376,5	normal	sobre	1	SI
87593	381,5	396,7	281,1	322,6	sobre	sobre	0	SI

Tabla 10. Pronóstico para el año 2016 estaciones meteorológicas de la Región 1.

La tabla 10. Muestra la verificación en la región 1, esta región corresponde a la región de CABA y alrededores, se puede observar que el porcentaje de acierto es del 100% en esta región (cinco sobre cinco estaciones).

idOMM	Observado	Pronóstico	Tercil Inferior	Tercil Superior	Clase.OBS	Clase.PRON	Diferencia	Acierto
87532	292,5	252,5	291,3	339,8	normal	sub	1	SI
87544	378,2	244,2	282,1	332,1	sobre	sub	2	NO
87550	360,7	248,9	313,5	349,4	sobre	sub	2	NO
87640	631,9	230,6	279,9	342,1	sobre	sub	2	NO

Tabla 11. Pronóstico para el año 2016 estaciones meteorológicas de la región 2.

La tabla 11. Muestra la verificación en la región 2, esta región corresponde al centro oeste de la provincia de Buenos Aires cercana a la Pampa, en esta zona el modelo tuvo un porcentaje de acierto fue de un 25% (una sobre 4 estaciones).

idOMM	Observado	Pronóstico	Tercil Inferior	Tercil Superior	Clase.OBS	Clase.PRON	Diferencia	Acierto
87244	307,0	339,4	341,8	382,1	sub	sub	0	SI
87328	238,0	366,8	306,6	352,3	sub	sobre	2	NO
87344	399,1	424,2	399,8	448,7	sub	normal	1	SI
87345	416,0	430,7	350,9	401,1	sobre	sobre	0	SI
87349	330,2	357,0	327,3	359,1	normal	normal	0	SI
87436	269,2	362,4	307,5	351,9	sub	sobre	2	NO
87448	273,0	313,9	302,5	324,2	sub	normal	1	SI
87453	242,4	384,2	337,2	401,6	sub	normal	1	SI

Tabla 12. Pronóstico para el año 2016 estaciones meteorológicas de la región 3.

La tabla 12. Muestra la verificación en la región 3, esta región corresponde a la zona más occidental de la región Pampeana, es una zona con orografía interesante, en esta zona el modelo regional tuvo un porcentaje de acierto regional del 75% (6 sobre 8 estaciones).

idOMM	Observado	Pronóstico	Tercil Inferior	Tercil Superior	Clase.OBS	Clase.PRON	Diferencia	Acierto
87563	440,3	279,0	233,5	310,4	sobre	normal	1	SI
87637	264,0	302,0	239,5	288,8	normal	sobre	1	SI
87641	265,4	299,4	267,0	311,1	sub	normal	1	SI
87645	242,6	299,0	260,0	297,6	sub	sobre	2	NO
87648	259,5	284,0	254,7	303,9	normal	normal	0	SI
87679	284,2	281,3	238,9	290,1	normal	normal	0	SI
87688	207,0	300,9	238,9	269,8	sub	sobre	2	NO
87692	351,9	269,7	251,8	283,9	sobre	normal	1	SI
87750	130,2	215,9	171,5	213,9	sub	sobre	2	NO

Tabla 13. Pronóstico para el año 2016 estaciones meteorológicas de la región 4.

La tabla 13. Muestra la verificación en la región 4, esta región corresponde a la zona sur de la provincia de Buenos Aires occidental de la región Pampeana, es la zona cuya media de precipitaciones es la menor de la región considerada, en esta zona el modelo regional tuvo un porcentaje de acierto regional del 67% (6 sobre 9 estaciones).

idOMM	Observado	Pronóstico	Tercil Inferior	Tercil Superior	Clase.OBS	Clase.PRON	Diferencia	Acierto
87371	585,0	299,3	274,6	333,7	sobre	normal	1	SI
87374	583,6	290,0	327,6	404,2	sobre	sub	2	NO
87467	564,9	271,4	348,2	390,2	sobre	sub	2	NO
87480	452,5	264,7	331,7	374,6	sobre	sub	2	NO
87497	615,2	241,1	303,4	374,3	sobre	sub	2	NO
87534	559,1	253,1	299,6	341,6	sobre	sub	2	NO
87548	484,0	252,2	329,4	373,3	sobre	sub	2	NO

Tabla 14. Pronóstico para el año 2016 estaciones meteorológicas de la región 5.

La tabla 14. Muestra la verificación en la región 5, esta región corresponde a la zona central de la región Pampeana que comprende la zona núcleo, se observa que el modelo para la región no pudo pronosticar las lluvias en exceso para el periodo de 2016. Quizás el modelo no consideró algún forzante climático importante regional (es algo para continuar investigando), en esta zona el modelo regional tuvo un porcentaje de acierto regional del 14% (1 sobre 7 estaciones).

idOMM	Observado	Pronóstico	Tercil Inferior	Tercil Superior	Clase.OBS	Clase.PRON	Diferencia	Acierto
87270	436,2	516,6	364,3	458,1	normal	sobre	1	SI
87289	440,9	535,4	356,9	440,7	sobre	sobre	0	SI
87393	448,0	487,8	405,4	475,8	normal	sobre	1	SI
87395	455,3	482,4	327,0	423,8	sobre	sobre	0	SI

Tabla 15. Pronóstico para el año 2016 estaciones meteorológicas de la región 6.

La tabla 15. Muestra la verificación en la región 6, esta región corresponde a la zona norte de la región Pampeana que comprende la zona núcleo, se puede observar que el porcentaje de acierto es del 100% (cuatro sobre cuatro estaciones).

El resumen de la verificación sobre las distintas regiones se muestra en la tabla siguiente:

Region	Aciertos	Total	Porcentaje de acierto
R1	5	5	100%
R2	1	4	25%
R3	6	8	75%
R4	6	9	67%
R5	1	7	14%
R6	4	4	100%
Total	23	37	62%

Tabla 16. Resumen de la verificación por Región

La figura 9 muestra el resultado regional de la verificación de los seis modelos regionales aplicados a cada una de las estaciones meteorológicas, se observa que en la región central de la región Pampeana los modelos no ajustaron bien, no pudiendo generar una clasificación aceptable. El resultado general fue de un nivel de acierto del 62% (23 estaciones sobre el total de las 37 estaciones). Los puntos verdes son las estaciones con acierto y los rojos en las que los modelos no ajustaron a las observaciones. Las regiones están identificadas con distintos símbolos, se los puede ver en la referencia de la figura, los símbolos fueron pintados de verde o rojo de acuerdo al acierto en el pronóstico o no.

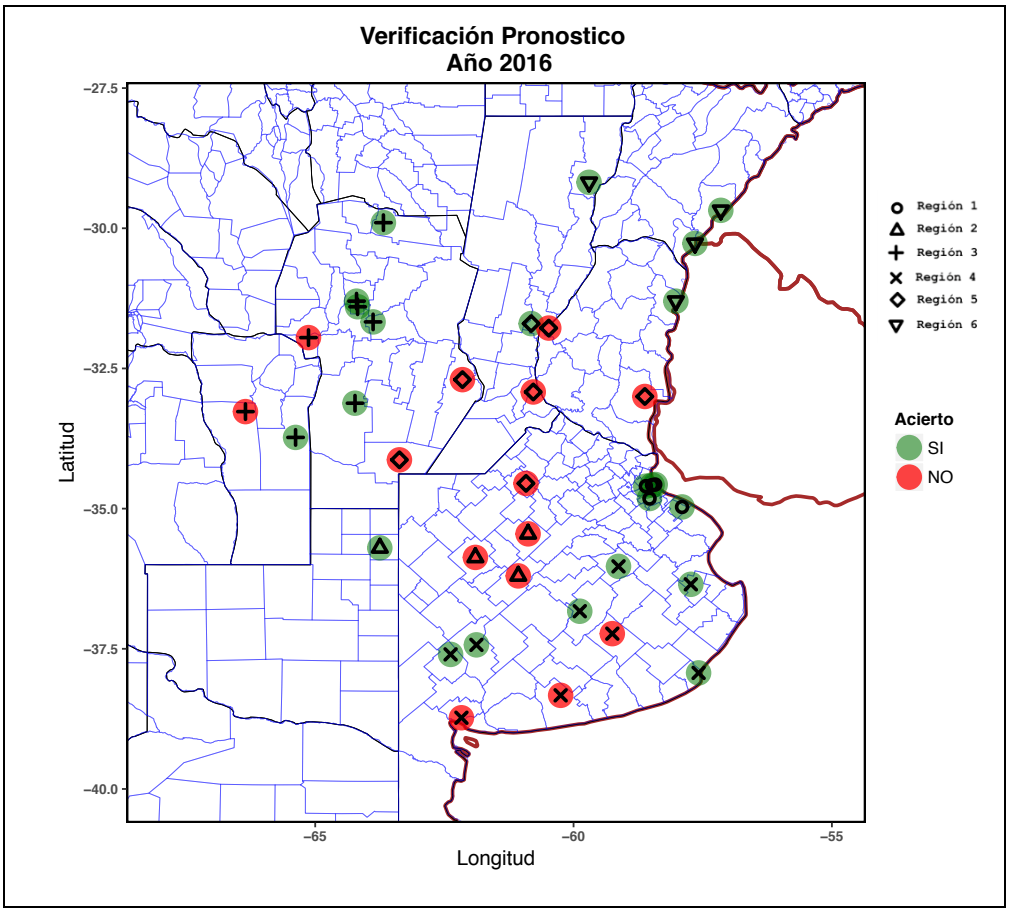


Figura 9. Verificación del pronóstico para el año 2016 (tasa de acierto del 62%)

4. Conclusión.

Es importante iniciar la conclusión de este trabajo reconociendo que la componente caótica asociada al comportamiento de la atmósfera es por naturaleza impredecible en escalas de tiempo estacionales e interanuales (Lorenz, 1996). De todos modos, en este trabajo se intentó evaluar la predictibilidad potencial sobre la región pampeana utilizando algunas metodologías estadísticas actuales. El objetivo de este estudio fue intentar construir una herramienta adicional al momento de la toma de decisión en un problema que involucre la precipitación estacional del próximo verano de la región pampeana.

Para ello se hizo un análisis exploratorio sobre la región pampeana de las observaciones provenientes de estaciones meteorológicas, con el objetivo de realizar una regionalización de las mismas, discriminando los efectos observados de los distintos regímenes de precipitación areal. Se probaron distintas técnicas disponibles actualmente para realizar el agrupamiento jerárquico y no jerárquico de las estaciones meteorológicas distribuidas no uniformemente sobre la región de estudio, comparándolas con metodologías actualmente en uso en el área de la meteorología, por ejemplo, el método de Lund. De este análisis, surgió que la mejor métrica para agrupar en la región pampeana fue la técnica de K-Means que es un método no jerárquico. Por otro lado, se obtuvieron buenos resultados usando un método jerárquico utilizando una métrica de distancia euclidea con un método aglomerativo de minimización de la varianza dentro del grupo (Ward.D). Para la validación de estos agrupamientos se aplicaron técnicas de validación interna de los grupos como silhouette, y técnicas de verificación de cantidad óptima de grupos. El resultado de los agrupamientos fue coherente con la variabilidad de la precipitación climatológica media en la región obteniéndose seis agrupamientos de estaciones meteorológicas. Además, se realizó una verificación usando análisis de la varianza (ANOVA), para evaluar si los grupos eran significativamente distintos, con resultados positivos. De este modo, se determinó que los grupos seleccionados fueron consistentes en la región de estudio. Con estos grupos se construyeron series medias regionales de precipitación.

Luego se construyeron mapas de correlación desfasada de los forzantes climáticos del mes previo al verano (noviembre) con las series medias de precipitación de verano de los agrupamientos de estaciones meteorológicas. Se encontraron, de este modo, regiones con correlación significativa sobre áreas interesantes desde el punto de vista de los forzantes climáticos. Se extrajeron las series medias de esas regiones que mostraron correlación significativa (mayor a 0.35 en valor absoluto). Se aplicó una técnica distinta a la tradicional en esta etapa para la generación de las series de predictores, usando los puntos interiores a las isolíneas del umbral de correlación significativa y no regiones rectangulares, que es la técnica usual. Se obtuvieron entre 15 y 40 predictores por cada subregión.

Se aplicó la técnica de LASSO para filtrar predictores poco significativos y evitar e problema de la colinealidad entre predictores. Con los predictores resultantes se construyó un modelo de regresión lineal múltiple para cada una de las 6 subregiones de estudio utilizando la técnica de validación cruzada en particular usando la técnica de "Leave-one-out cross-validation". Debido a que este tipo de modelos, como se verificó observando las figuras de las series de observación versus modelo, no pueden reproducir exactamente la variabilidad de la precipitación, en particular los extremos, se

categorizó a la precipitación en Sub-Normal, Normal y Sobre-Normal en base a los terciles de las series medias de precipitación de cada región.

Los modelos regionales obtenidos se aplicaron a cada una de las 37 series de precipitación de verano de la región para el año 2016, aplicando a cada serie el modelo derivado para la región a la cual pertenece. Es interesante destacar que el verano seleccionado de 2016 fue un año muy particular ya que fue el año Niño más extremo registrado desde el año 1950 y quizás sea la causa que en particular en la zona núcleo, la tasa de acierto del modelo fuera tan baja. Se puede ver que el modelo pronosticó precipitaciones un 50% más bajas que las observaciones. De todos modos, los modelos funcionaron aceptablemente bien en las subregiones, obteniendo una tasa de acierto regional del 62% a nivel regional.

Todos los programas y los productos de datos, así como los gráficos de este trabajo se pueden encontrar en:

https://github.com/alrolla/Especializacion_2018

5. Agradecimientos.

Al Servicio Meteorológico Nacional por la generosa provisión de los datos de precipitación.
Al National Centers for Environmental Prediction (NCEP) y al National Center for Atmospheric Research (NCAR), por su tarea diaria operativa de mantener en forma abierta sus datasets de variable grilladas climáticas.
A la Dra. Marcela Hebe González y su grupo por su apoyo y su tiempo explicándome y ayudándome a solucionar los problemas que afronté con el pronóstico estadístico estacional.

6. Referencias.

Ashok, K., Behera, S., Rao, S., Weng, H., Yamagata, T., 2007. El Nino Modoki and its possible teleconnection. J. Geophys. Res., 112, C11007, doi: 10.1029/2006JC003798.

Barros V., Doyle M. y Camilloni, I., 2008. Precipitation trends in southeastern South America: relationship with ENSO phases and the low-level circulation. Theoretical and Appl. Climatology 93, 1-2, 19-33.

Barreiro, M., 2010. Influence of ENSO and The South Atlantic Ocean on climate predictability over Southeastern South America. ClimDyn, 35, 1493-1508.

Bivand Roger,2018. Package ‘maptools’. Tools for Reading and Handling Spatial Objects.

Bivand Roger, 2018. Package ‘rgdal’. Bindings for the 'Geospatial' Data Abstraction Library.

Chan, S.C., Behera, S.K., Yamagata, T., 2008. Indian Ocean Dipole influence on South American rainfall, Geophys. Res. Lett.,35, L14S12, doi: 10.1029/2008GL034204.

Compagnucci R. and W. Vargas, 1998. Inter-annual variability of the Cuyo river streamflow in the Argentinean andean mountains and ENSO events. Int. J. Climatol. 18, 1593-1609.

de Vries Andrie, 2016. Package ‘ggdendro’. Create Dendrograms and Tree Diagrams Using 'ggplot2'.

Kiladis, G.N., Diaz, H.F., 1989. Global climatic anomalies associated with extremes in the Southern Oscillation. J. Climate 2, 1069–1090.

Garbarini E., Skansi M., Gonzalez M.H. and Rolla A., 2016. ENSO Influence over Precipitation in Argentina, Advances in Environmental Research, Chapter 7, Volume 52, NOVA Publisher, NY, USA.

González, M.H., Domínguez, D., Núñez, M., 2012a. Long term and interannual rainfall variability in Argentinean Chaco plain region, Rainfall: Behavior, Forecasting and Distribution, Cap. 4876, Nova Science Publishers Inc.

Grimm, A, Barros, V and Doyle, M., 2002. Climate variability in Southern South America associated with El Niño and La Niña events. J. Climate, 13, 35-58.

Hastie Trevor, 2018. Package 'glmnet'. Lasso and Elastic-Net Regularized Generalized Linear Models.

Hijmans Robert J., 2017. Package 'raster' Geographic Data Analysis and Modeling.

Hijmans Robert J., 2017. Package 'fpp'. Data for ``Forecasting: principles and practice".

Kalnay E, Kanamitsu, M. ,Kistler, R., Collins, W. , Deaven, D., Gandin, L., Iredell, M., Saha, S.,White, G., Woollen, J., Zhu, I., Chelliah, M., Ebisuzaki, W., Higgings, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D.,1996. The NCEP/NCAR Reanalysis 40 years- project. Bull Amer Meteor Soc, 77, 437-471.

Liebmann B., Vera, C., Carvalho, L., Camilloni, I., Hoerling, M., Allured, D., Barros, V., Báez, J. y Bidegain, M., 2004. An Observed Trend in Central South American Precipitation. J. Climate, 17, 22, 4357- 4367.

Lorenz, Edward (1996). "Predictability – A problem partly solved" (PDF). Seminar on Predictability, Vol. I, ECMWF.

Lund, I.A., 1963. Map pattern classification by statistical methods, Journal of Applied Met. 2, 56-65.

MacQueen (1967). Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967), 281--297.

Nan, S., Li, J. 2003. The relationship between summer precipitation in the Yangtse River Valley and the previous Southern Hemisphere Annular Mode, Geophys. Res. Lett., 30, 24, 2266.

Nnamchi, H.C., Li, J.P., Anyadike, R.N., 2011. Does a dipole mode really exist in the South Atlantic Ocean? J. Geophys. Res., 116, doi: 10.1029/2010JD015579.

Ooms Jeroen et al, 2018. Package 'RMySQL'. Database Interface and 'MySQL' Driver for R.

Pierce David, 2017. Package 'ncdf4'. Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files.

R Core Team, 2017. R: A Language and Environment for Statistical Computing.

Reboita M.S., AmbrizziT., Da Rocha R.P., 2009. Relationship between the Southern Annular Mode and Southern Hemisphere Atmospheric Systems. Revista Brasileira de Meteorología, 24, 1, 48-55.

Ropelewski, C., y Halpert, M., 1987. Global and Regional scale precipitation patterns associated with El Niño. Mon Wea Rev, 115, 8, 1606-1626. doi: http://dx.doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2

Rousseeuw, Peter J. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

Rousseeuw et al. 2018. Package ‘cluster’. Finding Groups in Data': Cluster Analysis Extended.

Saji, N.H., Goswami, B.N., Vinayachandran, P.N., Yamagata, T., 1999. A Dipole Mode in the tropical Indian Ocean. Nature, 401, 23,360–363.

Saurral, R., Inés A. Camilloni, Barros, V., 2016. Low-frequency variability and trends in centennial precipitation stations in southern South America, Int. J. Climatol. , Wiley Online Library, DOI: 10.1002/joc.4810

Steinhaus (1956). Sur la division des corps mat´eriels en parties. Bulletin de l’Académie Polonaise des Sciences, Classe III, vol. IV, no. 12, 801-804.

Stone M,1974. *Cross-validatory choice and assessment of statistical predictions.* J. Royal Stat. Soc., 36(2), 111–147.

Thompson, D.W. y Wallace, J.M., 2000. Annular modes in the extratropical circulation. Part I: month-to-month variability. J. Climate, 13, 1000-1016.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the lasso”. Journal of the Royal Statistical Society. Series B (methodological) 58 (1). Wiley: 267–88. http://www.jstor.org/stable/2346178

Vera, C., Silvestri, G., Barros, V. and Carril, A. 2004. Differences in El Niño response in Southern Hemisphere. J. Climate 17, 9: 1741-1753.

Walker Alexander, 2018. Package ‘xlsx’. Read, Write and Edit XLSX Files.

Wickham Hadley et al, 2017. Package ‘dplyr’. A Grammar of Data Manipulation.

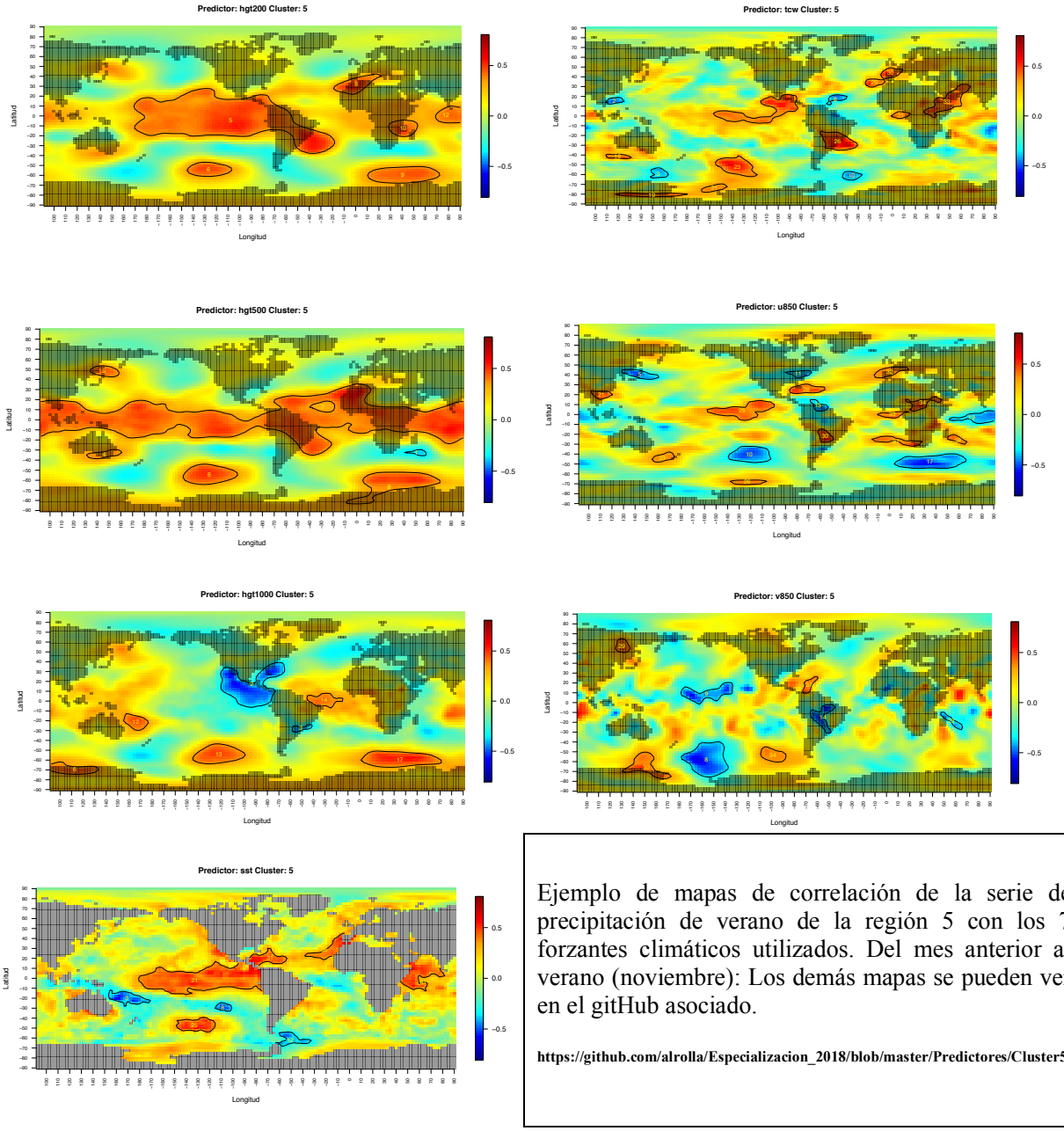
Wickham Hadley et al. 2018. Package ‘tydr’. Easily Tidy Data with 'spread()' and 'gather()' Functions.

Wickham Hadley et al. 2018. Package 'ggplot2'. Create Elegant Data Visualisations Using the Grammar of Graphics.

Wilks, D.S. Statistical methods in the atmospheric sciences (An introduction),1995. International Geophysics series. Academic Press, San Diego, California, USA, pp 467.

Wilks D. Improved statistical seasonal forecast using extended training data, 2008. Int. J Climatology, vol 28, 12, 1589-1598.

Figura 8. Mapas de correlación en la región 5 (Zona núcleo de la región Pampeana)



Ejemplo de mapas de correlación de la serie de precipitación de verano de la región 5 con los 7 forzantes climáticos utilizados. Del mes anterior al verano (noviembre): Los demás mapas se pueden ver en el gitHub asociado.

https://github.com/alrolla/Especializacion_2018/blob/master/Predictores/Cluster5/

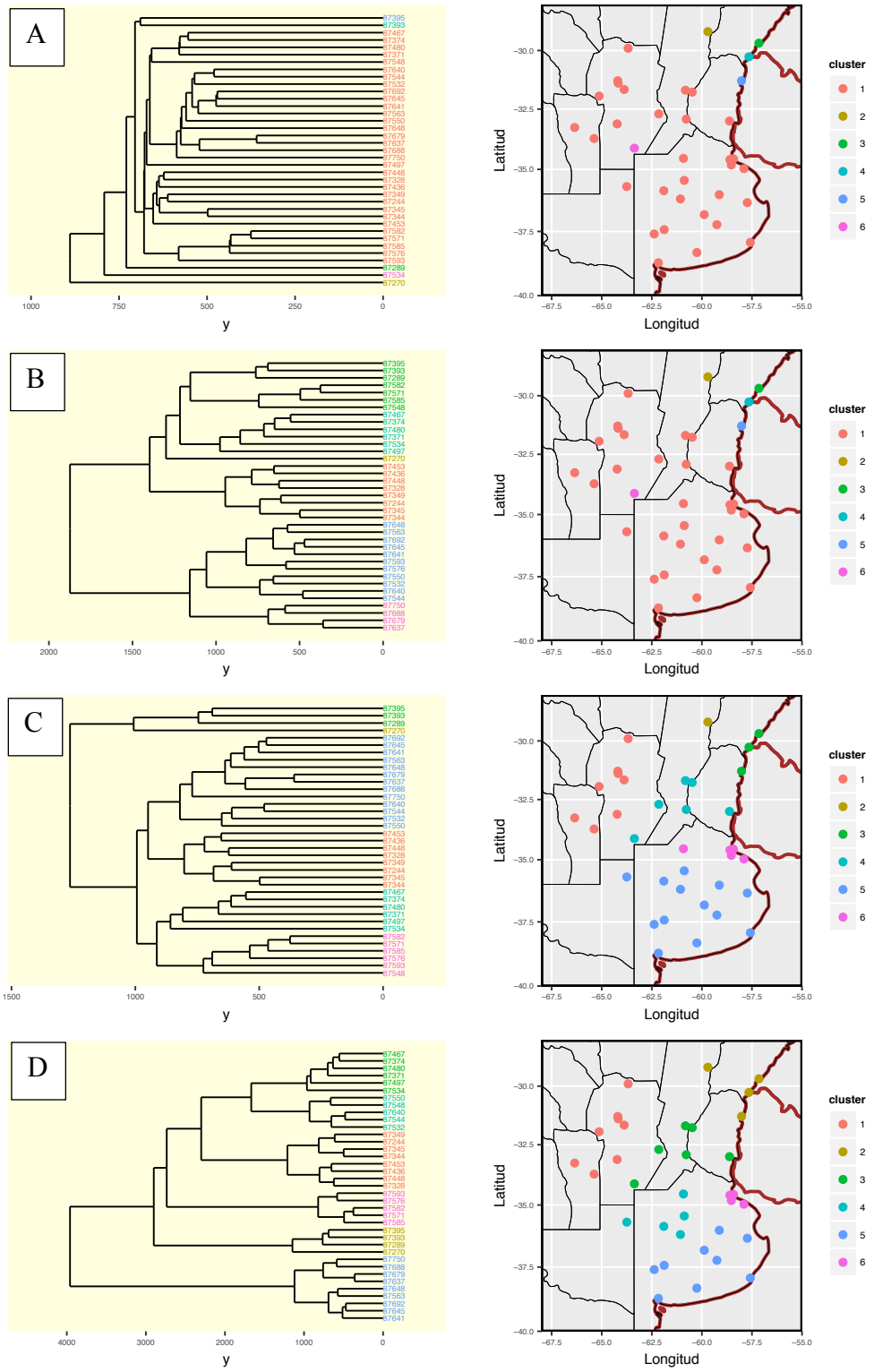
Tabla 1. Detalle de Estaciones Meteorológicas

Id	idOMM	Estación Meteorológica	Institución	Longitud	Latitud	Elev.
1	**87244**	Villa Maria del Rio Seco	SMN	-63.68	-29.90	341
2	**87270**	Reconquista Aero	SMN	-59.70	-29.18	53
3	**87289**	Paso de los Libres Aero	SMN	-57.15	-29.68	70
4	**87328**	Villa Dolores Aero	SMN	-65.13	-31.95	569
5	**87344**	Cordoba Aero	SMN	-64.20	-31.30	474
6	**87345**	Cordoba Obs.	SMN	-64.18	-31.40	425
7	**87349**	Pilar Obs.	SMN	-63.88	-31.67	338
8	**87371**	Sauce Viejo Aero	SMN	-60.82	-31.70	18
9	**87374**	Parana Aero	SMN	-60.48	-31.78	78
10	**87393**	Monte Caseros Aero	SMN	-57.65	-30.27	54
11	**87395**	Concordia Aero	SMN	-58.02	-31.30	38
12	**87436**	San Luis Aero	SMN	-66.35	-33.27	713
13	**87448**	Villa Reynolds Aero	SMN	-65.38	-33.73	486
14	**87453**	Rio Cuarto Aero	SMN	-64.23	-33.12	421
15	**87467**	Marcos Juarez Aero	SMN	-62.15	-32.70	114
16	**87480**	Rosario Aero	SMN	-60.78	-32.92	25
17	**87497**	Gualeguaychu Aero	SMN	-58.62	-33.00	21
18	**87532**	General Pico Aero	SMN	-63.75	-35.70	145
19	**87534**	Laboulaye Aero	SMN	-63.37	-34.13	137
20	**87544**	Pehuajo Aero	SMN	-61.90	-35.87	87
21	**87548**	Junin Aero	SMN	-60.92	-34.55	81
22	**87550**	Nueve de Julio	SMN	-60.88	-35.45	76
23	**87563**	Las Flores Aero	SMN	-59.13	-36.03	36
24	**87571**	El Palomar Aero	SMN	-58.60	-34.60	12
25	**87576**	Ezeiza Aero	SMN	-58.53	-34.82	20
26	**87582**	Aeroparque Buenos Aires	SMN	-58.42	-34.57	6
27	**87585**	Buenos Aires	SMN	-58.48	-34.58	25
28	**87593**	La Plata Aero	SMN	-57.90	-34.97	19
29	**87637**	Coronel Suarez Aero	SMN	-61.88	-37.43	233
30	**87640**	Bolivar Aero	SMN	-61.07	-36.20	94
31	**87641**	Azul Aero	SMN	-59.88	-36.83	147
32	**87645**	Tandil Aero	SMN	-59.25	-37.23	175
33	**87648**	Dolores Aero	SMN	-57.73	-36.35	9
34	**87679**	Pigue Aero	SMN	-62.38	-37.60	304
35	**87688**	Tres Arroyos	SMN	-60.25	-38.33	115
36	**87692**	Mar del Plata Aero	SMN	-57.58	-37.93	21
37	**87750**	Bahia Blanca	SMN	-62.17	-38.73	83

Tabla 2. Series medias anuales de precipitación regionales de verano.

Id	Año	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
1	**1979**	226,5	276,7	340,5	204,7	240,5	274,1
2	**1980**	325,1	405,1	504,3	288	443,3	545,5
3	**1981**	305,9	293,5	264,3	204,7	333,9	365,4
4	**1982**	209,4	327,5	400,8	244,7	300,3	462,8
5	**1983**	553,9	668	389,9	402,4	600,8	549
6	**1984**	240,8	234,5	425,9	246,3	212,1	273,1
7	**1985**	187,4	384,4	328,5	273,1	310,5	317,5
8	**1986**	261,7	374,8	353,1	173,9	286,9	264,1
9	**1987**	251,4	203,7	383,6	254,5	342,8	418,4
10	**1988**	155,3	194,8	294,1	160,6	208,5	181,4
11	**1989**	632,1	305,6	398,6	389,2	401,2	474,8
12	**1990**	309,7	229,2	380,5	314,8	374,3	376,4
13	**1991**	370,3	409,3	387,4	382,7	403,9	501,4
14	**1992**	328	409,6	375,7	230,6	405,8	418,4
15	**1993**	273,9	275,6	333,6	231,1	262,4	451,7
16	**1994**	243,1	331	324,9	259	333,1	365,8
17	**1995**	251,7	293,6	343,1	224,7	280,4	396,9
18	**1996**	302,2	402,5	389,2	362,9	289,8	426,9
19	**1997**	508,3	466,4	486,5	318,1	553,3	996,3
20	**1998**	491,8	239,5	239,8	215,7	301,5	392
21	**1999**	209,9	408,7	413	349,7	344,4	280,7
22	**2000**	418,2	316,4	326	258,4	363,9	474,4
23	**2001**	300	202,3	323,3	222	225	245,5
24	**2002**	568,4	274,3	314,9	238,1	428,1	527
25	**2003**	335,1	269,1	314,1	237,5	246,7	258
26	**2004**	278,8	444,1	373,2	323,7	395,8	347,2
27	**2005**	436,6	324,6	327,9	330,2	274,6	245,3
28	**2006**	384,5	353	413,9	330,2	539	588,1
29	**2007**	260,2	302,1	407,7	215,1	324	309,4
30	**2008**	175,7	137,2	268,3	117,4	279,3	291
31	**2009**	574,4	392,5	350,6	321,7	717,7	687,1
32	**2010**	280,4	384,4	314,4	277,2	349,7	430,5
33	**2011**	262,5	422	301,2	224,4	337,3	359,5
34	**2012**	386,2	238,8	254,4	292,8	352,4	611,2
35	**2013**	513,1	264,1	380,5	231,8	435,7	457
36	**2014**	272,9	193,9	547,3	249,7	455,9	737,3
37	**2015**	259,2	412,7	446,6	345,3	494,8	601,3
38	**2016**	356,8	415,8	309,4	271,7	549,2	445,1
39	**2017**	201,3	161,4	277,3	174,9	220,6	194,9

ANEXO I. Experimentos de agrupamiento jerarquico



Ejemplo de agrupamiento usando distancia euclidea, seleccionando 6 clusters, con agrupamiento :

A) Single linkage
B) Complete Linkage
C) Average
D) ward.D

El conjunto total de pruebas se pueden visualizar en :

https://github.com/alrolla/Especializacion_2018/tree/master/Agrupamientos