Universidad Nacional Autónoma de México

Facultad de Ciencias

Proyecto Final: selección genómica de potenciales híbridos de arroz de la variedad japónica (*Oryza sativa L.*). Una comparación entre el uso de regresión por componentes principales y el uso de bosques aleatorios.

Profesor: M.C. Sergio Hernádez López
Ayudantes:
Biól. Rafael López Martínez
M.C. Jazmín de Jesús Santillán Manjarrez

Equipo:
Cano Paez Bernabe
Martínez Aguirre Sharon
Sierra Casiano Yuznhio

6 de Junio de 2023

1. Introducción

La **Selección Genómica** es una metodología que ha revolucionado el fitomejoramiento, también conocido como "mejoramiento vegetal", que es la técnica para diseñar la genética de las plantas en beneficio de la humanidad (Poehlman, J.M. y Sleper, D.A. 1995); esto con el objetivo seleccionar a los individuos con el fenotipo deseado usando información previa (un conjunto de entrenamiento) que contiene información fenotípica y genotípica, con la que se ajusta y se evalúa un modelo estadístico o de aprendizaje automatizado.

El éxito de la Selección Genómica ha sido catalizado principalmente por la reducción considerable del tiempo de realización de los experimentos convencionales del fitomejoramiento y por la reducción significativa en el costo de tecnologías de genotipado, como lo es en el caso de la codificación por marcadores genéticos (Montesinos-López et al, 2022). Un **marcador genético** es un gen o posición en el genoma que existe en dos o más alelos distinguibles, con una localización conocida y cuya herencia puede ser, por lo tanto, seguida a través de un cruce genético, permitiendo mapear la posición de un gen a determinar (Sevilla, S. 2007). Es generalmente usado para identificar individuos, por lo que es una herramienta muy poderosa para explorar la diversidad genética en una población.

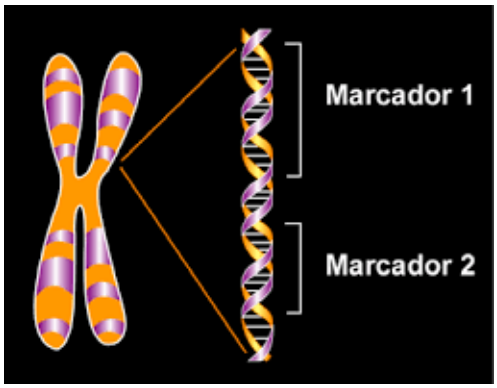

Figura 1: Representación esquemática de los marcadores genéticos.

Los SNP's (por sus siglas en inglés, Polimorfismo de Nucleótido Único: Adenina, Citocina, Timina y Guanina) son marcadores genéticos que consisten en cambios de un único nucleótido (un par de bases) en la secuencia de ADN en los alelos, los SNP puden ser polimorfismos bi, tri o tetraalélicos. Puden encontrarse en regiones codificantes, no codificantes o en regiones intergénicas y dependiendo de la localización producir mutaciones silenciosas, sin sentido, de cambio de sentido o afectar la unión de factores de transcripción al promotor o el splicing del ARN.

Un ejemplo de la diversidad o variación genética en una población diploide, en donde los individuos de la población poseen dos pares de cromosomas con dos alelos, A y a, es la proporción o frecuencia de pares de SNP's en la secuencia de ADN de los alelos, tenemos diferentes fecuencias alélicas para los tres genotipos (AA, Aa y aa), donde el *alelo menor* es el que corresponde al alelo con menor frecuencia y el otro es conocido como *alelo mayor*.

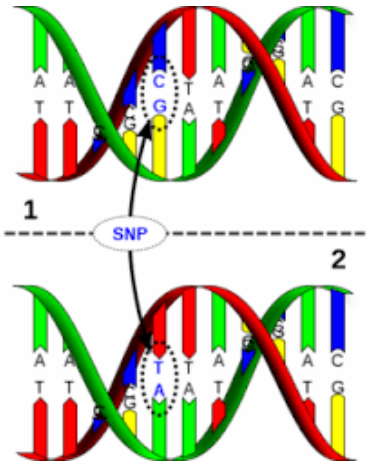

Figura 2: Representación esquemática de los pares de SNP's en los alelos.

Una vez que tenemos esta información sobre los alelos y los SNP's, se recodifica la información de los marcadores con base en las siguientes reglas:

$$x = \begin{cases} 0 & \text{si el SNP es homocigoto para el alelo mayor} \\ 1 & \text{si el SNP es heterocigoto} \\ 2 & \text{si el SNP es homocigoto para el alelo menor} \end{cases} \quad (1)$$

De esta forma, el objetivo de la Selección Genómica es predecir el fenotipo de los individuos de una población (híbridos) en ciertas condiciones ambientales y en función de las variables genotípicas para evaluar su rendimiento y así seleccionar a aquellos híbridos con características deseables.

Sin embargo, debido a que no existe un modelo universal, es necesario evaluar algunos modelos estadísticos o de aprendizaje automatizado para un conjunto de datos en particular y subsecuentemente elegir la mejor opción entre éstos, para predecir el fenotipo de nuevos individuos dentro del conjunto de datos.

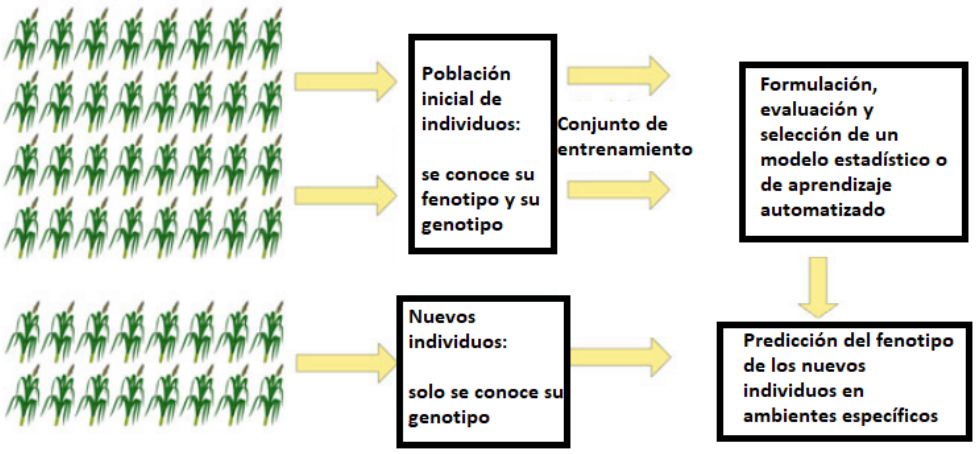

Figura 3: Representación esquemática de los elemento necesarios para implementar la selección genómica.

Planteamiento del problema

Monteverde et al. (2019) investigaron un conjunto de datos referente a una población de arroz tropical llamada **Japonica**, en el cual se consideran las cuatro variables fenotípicas de interés siguientes:

- GY: producción de grano de arroz en kilogramos por hectárea (Grain Yield).
- PH: altura promedio de la planta (Plant Height), desde el nivel del suelo a la punta de la hoja más alta.
- PHR: porcentaje de arroz recuperado, es decir, el porcentarje de arroz no desechado después de ser molido para perder su cáscara (Percentage of Head rice Recovery).
- GC: porcentaje de granos cálcareo, es decir, de granos que contienen calcio (percentage of Chalcky Grain).

La recolección de estos datos fue a lo largo de 5 años (2009, 2010, 2011, 2012 y 2013), que pueden ser considerados como ambientes distintos entre sí.

Figura 4: "El arroz *índica* y el arroz *japónica* son dos subespecies principales del arroz cultivado en Asia. El arroz *índica* se cultiva principalmente en entornos tropicales y subtropicales a latitudes o altitudes más bajas, mientras que el arroz *japónica* se cultiva principalmente en entornos más templados a latitudes o altitudes más altas." (Yang et al., 2014).

```
# Primeros renglones del fenotipo
> head(Pheno)
    Env     Line       GY      PHR       GC      PH
8  2009 Line_10 6283.084 0.610651 0.020509 79.5059
9  2009 Line_11 6551.095 0.639734 0.036922 79.1496
10 2009 Line_12 7270.608 0.578944 0.033119 88.8718
11 2009 Line_13 6693.288 0.618688 0.041221 93.5668
12 2009 Line_14 6404.684 0.591171 0.031192 86.6403
13 2009 Line_15 7219.528 0.604634 0.040226 98.9987
```

En este conjunto de datos, un total de 320 híbridos distintos fueron cultivados a lo largo de los 5 años. La siguiente tabla muestra la cantidad de híbridos que cuentan con su fenotipo en el conjunto de datos Japonica en cada año:

Año	Cantidad de híbridos que cuentan con su fenotipo
2009	93
2010	292
2011	316
2012	316
2013	134
Total	1151

Resulta natural preguntarse por el rendimiento de cada híbrido a lo largo de los 5 años; sin embargo, la tabla anterior nos muestra que no hay un solo año en el que se cuente con los fenotipos de cada uno de los 320 hibrídos.

El **objetivo principal** de este proyecto final es comparar los métodos de Regresión por Componentes Principales y Bosques Aleatorios y utlizar el mejor método para predecir los datos no observables en el conjunto de datos (el fenotipo de las hibrídos no observados a lo largo de los años) para proponer algunos híbridos con un rendimiento potencialmente mayor al resto.

Además, por simplicidad, **decidimos trabajar unicamente con la variable GY** como fenotipo ya que ésta nos parece la variable fenotípica de mayor interés, al medir directamente el rendimiento de la producción de grano de arroz en kilogramos por hectárea.

Modelo Planteado

El modelo estadístico planteado es el usado por Montesinos-López et al. (2022):

$$Y_{ij} = \mu + E_i + g_j + (gE)_{ij} + \varepsilon_{ij}$$

donde Y_{ij} es la producción de granos de arroz (Kg/ha) del j-ésimo híbrido en el i-ésimo ambiente; μ es la producción media de granos de todos los híbridos en todos los ambientes; E_i es el efecto del i-ésimo ambiente; g_j es el efecto del j-ésimo híbrido; $(gE)_{ij}$ es el efecto de la interacción fenotipo-ambiental; y ε_{ij} es el componente del error aleatorio del modelo.

Todo esto considerando que $E1$, E_2, E_3, E_4, y E_5 son lo efectos de los ambientes de los años 2009, 2010, 2011, 2012 y 2013, respectivamente. Además, $j \in \{1, 2, \cdots, 320\}$.

Cabe mencionar que, dado que se busca predecir datos no observables dentro de cada uno de los ambientes y no en nuevos ambientes, no se incluyeron covariables ambientales como precipitación media, temperatura promedio, radiación solar, etc., pues el efecto de estas variables latentes (o no observables) es considerado conjuntamente en el efecto de cada ambiente (E_i).

Pre-procesamiento de los datos

Previamente observamos las primeras observaciones de las variables fenotípicas, identificadas por los híbridos (*Line*) cultivados en alguno de los ambientes (*Env*). Sin embargo, observar las variables genotípicas resulta difícil, pues **el marcador secuenciado consta de 16,383 SNP's**, como a continuación se verifica y se muestran los primeros 6 renglones de las primeras 6 columnas

```
# Marcadores: cada renglon corresponde a la codificacion de los SNP's
#                para cada uno de los 320 hibridos
> dim(Markers)
[1]   320 16383

# Primeras codificaciones de los SNP's para los primeros hibridos
> head(Markers[, 1:6])
         S1_304168 S1_538661 S1_538663 S1_538674 S1_594195 S1_689889
Line_1           0         0         0         0         0         0
Line_10          1         0         0         0         2         0
Line_100         0         0         0         0         0         0
Line_101         0         0         0         0         0         0
Line_102         0         0         0         0         0         0
Line_103         0         0         0         0         0         0
```

VanRaden (2008) propone en su artículo "Efficient methods to compute genomic predictions", usar una matriz a la que llama "**matriz de relación genómica**", que se calcula como

$$\mathbf{G} = \frac{1}{p}\mathbf{X}\mathbf{X}^T$$

en donde **X** es la matriz de marcadores (cada renglón corresponde a la codificación del marcador para cada híbrido) y $p = 16,383$ corresponde al número de columnas (SNP's del marcador) de ésta matriz.

Notemos que ésta tranformación de la matriz de marcadores resulta bastante útil, pues reduce la dimensión de las 16,383 variables a 320 variables, coincidiendo además con aplicación de un kernel lineal a la matriz de marcadores.

Por otro lado, para considerar el efecto de cada ambiente (año) y de cada híbrido en su respectivo fenotipo (GY), éstos son considerados como varriables categoricas en modelo (indicando con 1 si la observación se realizó en el ambiente especificado y con 0 en caso contrario), de tal forma que se procesan mediante matrices diseño como a continuación se muestra:

```
> # Pre-procesamiento de los datos
> Line <- model.matrix(~0 + Line, data = Pheno)
> Env <- model.matrix(~0 + Env, data = Pheno)
> head(Env)
   Env2009 Env2010 Env2011 Env2012 Env2013
8        1       0       0       0       0
9        1       0       0       0       0
10       1       0       0       0       0
11       1       0       0       0       0
12       1       0       0       0       0
13       1       0       0       0       0
```

Finalmente, Montesinos-López et al. (2022), proponen usar la información de la matriz *Env*, incluir la información de la matriz de relación genómica colocando cada renglón de ésta correspondiente al híbrido de cada observacción (al multiplicar la matriz diseño de los híbridos *Line* por la matriz de relación genómica) e incluír una tercera matriz que uncluya la información genómica en interacción con la imformación ambiental, como se muestra a continuación:

```
> # Pre-procesamiento de los datos
> Line <- model.matrix(~0 + Line, data = Pheno)
> Env <- model.matrix(~0 + Env, data = Pheno)

> LineG <- Line %*% Geno
> LinexGenoxEnv <- model.matrix(~ 0 + LineG:Env)
>
> # Variables predictoras
> X <- cbind(Env, LineG, LinexGenoxEnv)
> dim(X)
[1] 1151 1925
>
> # Variable respuesta
> y <- Pheno$GY
```

de tal forma que **se cuenta con 1925 variables predictoras en un conjunto de 1151 observaciones**, siendo el caso en que la mayoría de éstas variables carece de interpretación individual, a excepción de las variables de la matriz *Env*.

Dado que se tienen más variables predictoras que observaciones, es razonable pensar que los modelos de regresión lineal sufren de multicolinealidad, motivando el uso de alguna técnica de reducción de dimensionalidad. Por esta razón, **proponemos el uso de Regresión por Componentes Principales** como alternativa al problema de multicolinealidad **y el uso de Bosques Aleatorios** como alternativa a este modelo estadístico multivariado.

2. Marco teórico

Para elegir el "mejor" modelo entre los dos modelos propuestos, elegimos usar el esquema de validación cruzada 5-fold CV para cada modelo y comparar el redimiento de cada modelo en términos del Error Cuadrático Medio (*MSE*).

Para esto, primero explicaremos en qué consiste este esquema de validación cruzada y posteriormente explicaremos los dos modelos propuestos, enfatizando en la Regresión por Componentes Principales, pero sustentada en el *Análisis de Componentes Principales*.

Finalmente, con los resultados obtenidos, discriminamos el mejor modelo en términos del MSE y con éste haremos las predicciones de los datos faltantes, con **el objetivo de proponer aquellos híbridos que mejor rendimiento promedio tuvieron a lo largo de los 5 años**.

▪ Validacion Cruzada

La Validación cruzada, al igual que la comparación de parámetros obtenidos, nos ayuda a validar los métodos de regresión garantizando que la partición de datos entre entrenamiento y prueba sea independiente. Es una manera de predecir el ajuste de un modelo a conjunto de datos de prueba cuando no disponemos de este.

Este método se basa en el cálculo de una media aritmética obtenida de las medias de cada partición.

La validación cruzada proviene del método *hold-out* el cual consiste en separar el conjunto de datos en dos subconjuntos usando uno directamente para entrenar el modelo y el segundo para realizar el test, de validación de forma que la función de aproximación solo se ajusta con los datos de entrenamiento y posteriormente calcular los valores de salida para el conjunto de datos de prueba.

El método *hold-out* es un método rápido al computar, por lo que podemos repetir el proceso tomando distintos conjuntos de datos de entrenamiento tomados aleatoriamente, así se calculan los estadísticos de regresión a partir de la media de cada una de las particiones creadas en cada repetición.

- *Validacíon Cruzada k-fold*

  Consiste en dividir los datos en k subconjuntos, es decir, se aplica el método *hold-out* k veces, así utilizamos un subconjunto de datos a la vez como datos de prueba y los $k-1$ grupos restantes como datos de entrenamiento. En este caso el error medio que obtenemos de los k análisis realizados nos arroja el error del método. Es decir

  $$E = \frac{1}{k}\sum_{i=1}^{k} E_i$$

  Este método tiene la ventaja de que todos los datos son considerados en al menos un subconjunto de datos, ya sea para validación o entrenamiento, por lo que los resultados se ajustan mejor a los datos.

▪ Regresión por componentes principales

- **Descomposición por Valores Singulares**

En este apartado, haremos mención de algunas definiciones y teoremas para justificar formalmente lo implementado para la regresión por componentes principales (PCR).

Dada una matriz $A \in R^{mxn}$, la matriz A^TA es simética, pues $(A^TA)^T = A^T(A^T)^T = A^TA$, y semidefinida positiva, ya que $x^TA^TAx = (Ax)^T(Ax) = \|Ax\|^2 \geq 0 \ \forall x \in \mathbb{R}^n$.

Con ello, se tiene que los eigenvalores de A^TA son reales y no negativos.

Definición 1: Sea $A \in R^{mxn}$. Sean $\lambda_1, \lambda_2, ..., \lambda_n$, los eigen valores de A^TA ordenados de forma decreciente. Entonces $\sigma_i = \sqrt{\lambda_i}$ es el i-ésimo valor singular de A

Teorema 1: Sea $A \in R^{mxn}$. Sean $\lambda_1, \lambda_2, ..., \lambda_n$, los eigen valores de A^TA y además,

$$\lambda_1, \geq \lambda_2, ..., \lambda_r > \lambda_{r+1} = ... = \lambda_n = 0$$

Es decir, los eigenvalores de A^TA están ordenados en forma decreciente y el número de eigenvalores no nulos es r. Sea $v_1, v_2, ...v_n$ una base ortonormal de $\mathbb{R}^n$ tal que $A^TAv_i = \lambda_i v_i$. Entonces.

1. $Av_1, Av_2, ...Av_n$ es un conjunto ortogonal y $\|Av_i\| = \sqrt{\lambda_i} = \sigma_i$ para todo $i = 1, 2, .., n$
2. $\frac{Av_1}{\sigma_1}, \frac{Av_2}{\sigma_2}, ...\frac{Av_r}{\sigma_r}$ es una base ortonormal para las columnas de A
3. $v_{r+1}, v_{r+2}, ...v_n$ es una base ortonormal para Nul(A)
4. rango(A) = r = número de valores singulares no nulos de A

A partir de lo anterior, definiremos lo que se conoce como descomposición en valores singulares de una matriz.

Definición 2: Sea $A \in R^{mxn}$. Una descomposición en valores singulares de A es una factorización

$$A = U\Sigma^T$$

con $U \in \mathbb{R}^{nxn}$ ortogonales y $\Sigma \in \mathbb{R}^{mxn}$ con,

$$\Sigma = \begin{bmatrix} D & 0_{rx(n-r)} \\ 0_{(m-r)xr} & 0_{(m-r)x(n-r)} \end{bmatrix} \qquad y \qquad D = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}$$

De donde $\sigma_1 \geq \sigma_2 \geq ...\sigma_r > 0$
De esta manera, si $A = U\Sigma V^T$ es una DVS de A, con Σ como en la definición 2, v_i y u_i son las i-ésimas columnas de V y U respectivamente, entonces A puede ser escrita de la forma

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$

Definición 3: Si $A = U\Sigma V^T$ es una DVS de A, los vectores que aparecen como columnas de la matriz V se les denomina *vectores singulares derechos de* A, mientras que a los que aparecen como columnass de U se les denomina *vectores singulares izquierdos de* A.

- **Explicación intuitiva**

Para la implementación de este método usamos la librería **pls**, la cuál es usada en diversos ámbitos científicos incluídas las ciencias naturales. La principal razón de su utilidad radica en la necesidad de reducir las variables predictoras con las que se está trabajando, principalmente porque es muy probable que muchas de ellas estén altamente correlacionadas.
Sabemos que la solución por mínimos cuadrados de la regresión lineal multivariada de la forma:

$$Y = XB + E$$

está dada por:

$$B = (X^T X)^{-1} X^T Y$$

Sin embargo, es usual en la práctica que la matriz $X^T X$ sea singular debido a que el número de variables en X excede el número de columnas; o bien también puede ser producto de la colinearidades. Es así que las funciones PCR y PLSR evitan lo anterior, descomponiendo la matriz X por valores singulares T y una matriz de cargas P, teniendo ahora que $X = TP$

De esta manera, la regresión no se aplicaría únicamente con X, adicionalmente se tomarían en cuenta las primeras columnas de. de T. En componentes principales, dichas columnas corresponden a los vectores singulares izquierdos de X multiplicados por su valor singular correspondiente. Además, las llamadas "cargas", son vectores singulares derechos de X. Con lo anterior, elegimos las cargas y scores óptimos para describir tanto como sea posible la covarianza entre X y Y, en donde la regresión por componentes principales se enfoca en explicar la varianza máxima de X

- **Algoritmo**

Como se mencionó anteriormente, aproximamos la matriz X a partir de los primeros a componentes principales, obtenidas de la descomposición de valores singulares (SVD). Así,

$$X = \tilde{X}_{(a)} + E_X = (U_a D_{(a)}) V_{(a)}^T + E_X = T_{(a)} P_{(a)}^T + E_X$$

.

Quedando entonces los coeficientes de regresión como sigue:

$$B = P(T^T T)^{-1} T^T Y = V D^{-1} U^T Y$$

Para PLSR, los componentes llamados variables latentes en este contexto, son obtenidos iterativamente. Se inicia aplicando la descomposición de valores singulares del producto cruz de la matriz $S = X^T Y$. De esta modo, se incluye información sobre la variación en X y Y y la correlación entre ellos. Los vectores singulares izquierdos y derechos w y q son usados como vectores de peso para X y Y respectivamente para obtener los scores t y u:

$$t = Xw = Ew$$
$$u = Yq = Fq$$

De donde E Y F se inicializan como X y Y respectivamente. Cabe mencionar que los scores t, son normalizados en algunas ecuaciones, i.e:

$$t = \frac{t}{\sqrt{t^T t}}$$

Por otro lado, las cargas para X y Y se obtienen como:

$$p = E^T t$$
$$q = F^T t$$

Finalmente las matrices de datos son "reducidas" de la siguiente forma: La información relativa a las variables latentes en la forma del producto exterior de tp^T y tq^T es sustraído de las matrices de datos E y F para cada iteración.

$$E_{n+1} = E_n - tp^T$$
$$F_{n+1} = F_n - tq^T$$

La estimación para el siguiente componente inicia a partir del SVD del producto cruz de $E_{n+1}^T F_{N+1}$. Después de cada iteración, los vectores w,t,p y q son almacenados como columnas en las matrices W, T,P y Q respectivamente. Sin embargo, existe una complicación para la matriz W, pues no pueden ser comparadas directamente a razón de que son generadas a partir de las matrices "reducidas" E y F. No obstante, podemos representar alternativamennte a los pesos de modo que las columnas se relacionen con la matriz original X como sigue:

$$R = W(P^T W)^{-1}$$

A partir de los procesos anteriores, ya estamos posicionados en el caso del PCR: En vez de aplicar la regresión a Y con X, usamos los "scores" T para calcular los coeficientes de la regresión para luego regresar a los correspondientes a la variable original premultiplicando con la matriz R, pues $T = XR$. Así:

$$B = R(T^T T)^{-1} T^T Y = RT^T Y = RQ^T$$

Siendo enfáticos en que sólo se usarán los primeros a componentes. Para saber cuál es el número óptimo se usará validación cruzada. Este método será analizado en el siguiente apartado.

- **Escogiendo el número de componentes con validación cruzada**

Tanto en la función **pls** y **pcr** se emplea por default éste método para la selección de variables, si se escribe "CV" como argumento, se aplicará validación cruzada usando diez particiones escogidas aleatoriamente (10-Fold Cross Validation), pero también podría ser de manera consecutiva o intercalada, según se desee modificando el argumento de *segment.type*.

Cuando se realiza dicha validación, el modelo contendrá métricas que nos permitirán conocer el poder predictivo, tales como el MSEP, R^2; o el MSE en nuestro caso. Siempre ha representado un reto la elección de cuántas variables se tiene que trabajar.

La función *selectNcomp* considera dos estrategias; una heurística, que consiste en elegir el modelo con menor cantidad de componentes que se encuentre a un error estándar del mejor modelo. La segunda estrategia es a partir de una selección de variables por pasos; ya sea agregando variables en cada iteración (forward) o retirando variables (backward) hasta llegar al modelo óptimo.

▪ Regresión por Bosques Aleatorios

- **Explicación intuitiva**

  La técnica de Bosques Aleatorios es una alteración de los árboles de decisión, basada conjuntamente en la técnica **bootstrap**, que toma una colección de árboles de decisión y promedia las respectivas predicciones de cada árbol en un conjunto de prueba. Para implementar esta técnica se consideró un criterio de partición de regiones basado en el error cuadrático medio (*MSE*), pues la variable respuesta es continua.

  Para entrenar o ajustar el modelo, se toman B muestras aleatorias del conjunto de entrenamiento (un hiperparámetro a calibrar) y se seleccionan aleatoriamente subconjuntos de tamaño $\lfloor p/3 \rfloor$ (donde $p = 1925$ es el número de variables predictoras) del conjunto de variables predictoras como predictores candidatos para cada una de éstas B muestras; todo esto, con el objetivo de ajustar un árbol de decisión para cada una de éstas muestras. Finalmente, el valor predicho de las observaciones del conjunto de prueba es calculado como el promedio de las predicciones de los árboles ajustados anteriormente; es decir, como

  $$\hat{y}_i = \frac{1}{B} \sum_{b=1}^{B} T_b(\ \mathbf{x}_i)$$

  donde T_b es el b-ésimo árbol de decisión ajustado.

- **Implementación del Algoritmo**

  Para cada partición identificada en el esquema de validación cruzada *5-fold CV* se identifica el conjunto de entrenamiento y el conjunto de prueba. Con el conjunto de entrenamiento se ajusta un Bosque Aleatorio con la función *random_forest()* de la librería *SKM* (un wrapper de la función *randomForestSRC::rfsrc()*), con los siguientes argumentos:

  - **trees_number**: se propusieron valores entre 50 árboles y 500 árboles.
  - **node_size**: se propusieron vaalores entre 5 y 15 como número de nodos terminales de cada árbol.
  - **tune_type**: se propuso usar optimización bayesiana como tipo de calibración de los hiperparámetros anteriores, para reducir la carga computacional del ajuste de éste modelo.
  - **...**: los demás argumentos se dejaron específicados por defecto.

3. Resultados

Una vez aplicados los dos métodos, obtuvimos el MSE global y para cada ambiente. A continuación, se muestra una tabla con los resultados mencionados.

MSE para Regresión por Componentes Principales			
Ambiente	MSE	MSE SD	MSE SE
2009	566210.8412	85563.0885	27057.4243
2010	827549.257	107446.477	33977.5594
2011	616666.607	61022.0241	19226.8641
2012	583671.115	84886.2736	26843.3967
2013	380175.655	50475.3202	15961.6977
Global	539054.652	60997.8834	19289.2244

MSE para Random Forest			
Ambiente	MSE	MSE SD	MSE SE
2009	591357.172	46720.00009	14774.1615
2010	642462.661	94250.6091	29804.6596
2011	552679.005	54155.5323	17125.483
2012	538491.805	84866.3828	26837.1066
2013	336148.954	48171.7984	15233.2602
Global	465782.917	51283.1186	16387.909

Con la información recabada, se presenta una gráfica para comparar el poder de predicción de cada modelo.

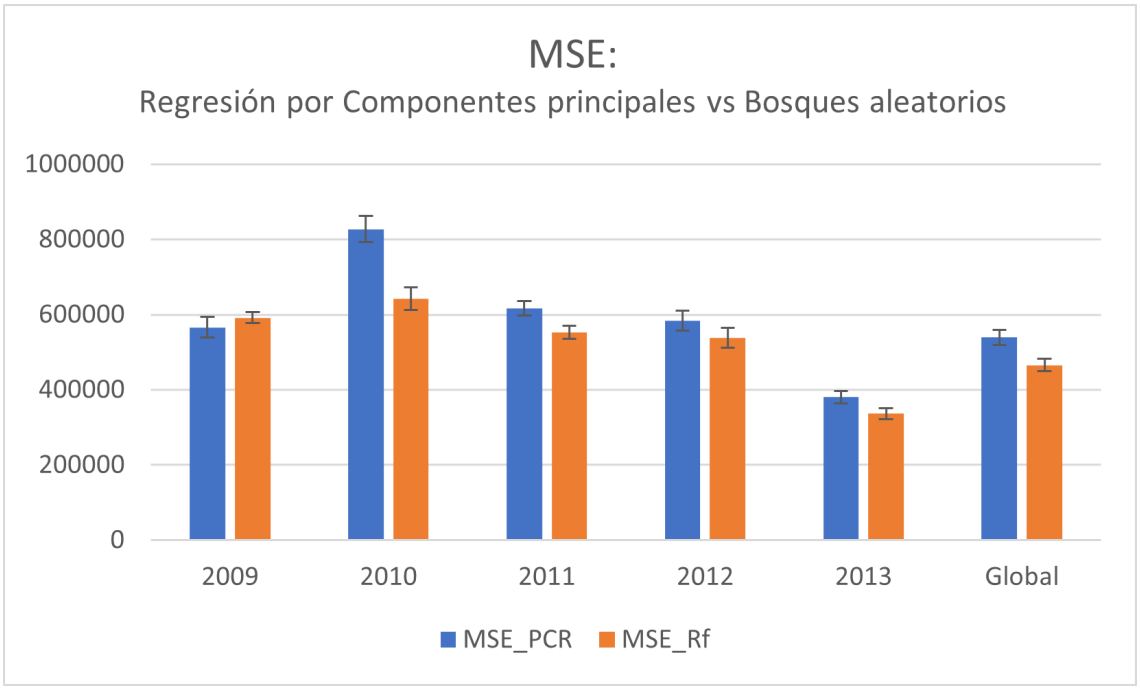

Figura 5: Comparación del poder predictivo de PCR y Random Forest

De manera global, pudimos observar como el MSE usando PCR es mayor comparado con la implementación de Random Forest. Además, se tiene que la longitud del intervalo de confianza para PCR es menor sugiriendo hasta ahora un mejor desempeño de Random Forest al tener una menor variabilidad.

Si analizamos el MSE para cada ambiente, tenemos un comportamiento similar descrito anteriormente, salvo el ambiente del 2009, ya que en este caso se tuvo un mayor puntaje en el MSE para random forest; teniendo el PCR una ganancia del 4.25

El ambiente de 2010 fue el que registró la mayor diferencia en el desempeño, pues se obtuvo una ganancia para random forest de 28.80 %. Aunque es importante mencionar que fue el ambiente en donde mayor MSE se obtuvo para ambos métodos. Será necesario investigar más a fondo las variables contempladas para ver cuáles están impactando negativamente en el desempeño del modelo.

En contraparte, vemos que para el ambiente de 2013 fue donde se registró el menor MSE para ambos métodos (180,175.65 para PCR y 336,148.95 para Random Forest). De hecho también fue el ambiente, cuyos intervalos de confianza fueron más pequeños; confirmando su buen desempeño.

Como observación general, la variabilidad en el MSE es mayor al usar PCR, pues independientemente de haber sido mejor que Random Forest para algún ambiente, los intervalos de confianza fueron más grandes.

3.0.1. Predicciones

Ahora, posterior a definir el modelo con *random_forest*, nos interesa el predecir los datos faltantes, por lo tanto desarrollamos:

```
> predictions <- predict(model, X_new)
> predictions
> predictions
$predicted
   [1]  6854.766  7000.312  7012.159  6984.496  7003.332  7057.799  6892.672
   [8]  7428.788  7312.983  7152.179  7142.659  7374.736  7469.251  7312.895
  [15]  7084.875  6848.109  7044.192  7166.596  6958.143  6883.941  6879.157
...
```

Al obtener nuestras predicciones, cambiamos los valores faltantes para GY

```
> NEWDAT<-mutate_at(New_Pheno, "GY", ~replace(., is.na(.), pred))
```

Sobre estos nuevos datos, agrupamos de acuerdo al tipo de híbrido, que en nuestros datos de encuentra como *Line* y calculamos el promedio de GY para cada híbrido.

```
> Promedios= Grupos%>% summarise(
+    Promedio = mean(as.numeric(GY)))
```

De estos promedios obtenemos los 10 mejores:

```
> final= Promedios[order(-Promedios$Promedio),]
> final[1:10,]
```

Line	Promedio
Line_165	10181
Line_230	10122
Line_249	9969
Line_309	9927
Line_280	9819
Line_159	9810
Line_247	9807
Line_229	9796
Line_251	9705
Line_131	9700

Así, los "mejores" híbridos son los que se encuentran en el cuadro anterior. Cabe recalcar que esta interpretación de "mejores híbridos" nos muestra aquellos que tienen un mayor rendimiento en producción de grano, sin considerar las demás características deseables en un "híbrido".

4. Discusión

Según Montesinos-López et al., la Selección Genómica es una metodologìa predictiva que ha revolucionado el fitomejoramiento, para la selección de individuos candidatos, gracias al uso de conjuntos de datos de referencia (conjuntos de entrenamiento), acelerando la razón de "ganacia genética" de cultivos como el maíz, el trigo, el garvanzo, el arroz, etc.; además de reducir los costos en el proceso convencional del fitomejoramiento, gracias a una reducción significativa en el costo de tecnologías de genotipado o secuenciación.

Sin embargo, hay muchos factores que necesitan ser considerados de manera cuidadosa para poder aplicar exitosamente la Selección Genómica. Algunos de estos factores son la complejidad biológica de algunos fenotipos que se desean predecir, la calidad de los datos de refenrencia (tanto las mediciones de los fenotipos como la secuenciación de los marcadores), el objetivo de predicción (por ejemplo, en este proyecto se ha predicho el fenotípo de híbridos que faltaron en algunos ambientes pero estaban presentes en otros, sin tener que predecirlos en nuevos ambientes) y la elección adecuada de algún modelo de aprendizaje estadístico automatizado.

La importancia de la calidad de los datos de referencia y de la elección adecuada de algún modelo de aprendizaje estadístico automatizado radica en el objetivo de poder entender las relaciones entre las variables de interés (en este caso, poder predecir el fenotipo de individuos en función del fenotípo, del ambiente y de su genotípo); es decir, la calidad de los resultados radica fuertemente en la calidad de los materiales y de la metodología. Además, muchos métodos y modelos han sido propuestos, comparados e implementados; como lo han sido pricipalmente los modelos lineales mixtos y sus versiones bayesianas, los métodos de regresón penalizada (Ridge y LASSO) y algoritmos de aprendizaje automatizado, como Bosques Aleatorios, Gradient Boosting Machine, redes neuronales, y regresión por mínimos cuadrados parciales.

No cabe duda de que se puede explorar la aplicación de nuevos métodos y modelos en Predicción genómica. Incluso se pueden explorar ligeras variantes de los modelos y métodos mencionados en el parráfo anterior.

Por otro lado, la complejidad biológica del fenotipo y de genotipo de un individuo nos lleva a creer que la aplicación de esta metología resulta ser "reduccionista". Sin embargo, "Todos los modelos son erróneos, pero algunos son útiles" (George E. P. Box), es decir, el problema en sí mismo no es sencillo de modelar, ya que una de las principales desventajas es el solo haber predicho una de las variables fenotípicas, ignorando las otras tres variables (la altura de la planta, el porcentaje de arroz recuperado, y el porcentaje de grano calcáreo), con lo que propusimos los diez "mejores" híbridos en cuenstión de cantidad (producción de granos en Kg/ha) pero ignorando la calidad de éstos, la cual puede ser regida por la altura de la planta o el porcentaje de grano calcáreo, entre otras, en ventaja de los productores y de los consumidores.

Otra limitante en el proyecto es el no haber incluído covariables ambientales (no lantentes), pues resulta natural creer que el fenotipo de un individuo despende fuertemente del medio en el que se desarolla, con el objetivo de predecir el fenotipo de los 320 híbridos en algún ambiente nuevo que puede ser caracterizado por éstas. Aunque el objetivo de predicción no fue este, el desarrollo del proyecto se limita a poder predecir el fenotipo de los híbridos faltantes de cada uno de los ambientes; es decir, solo podemos predecirlos dentro de

los ambientes pero no en nuevos ambientes (pues se cuenta con variables lantentes ambientales, pero no con las variables ambientales en sí).

Por otro lado, **¿Qué consecuencias tiene la Predicción y la Selección Genómica?** Es bien sabido que el fitomejoramiento es un proceso practicado por civilizaciones modernas y antiguas, como ha sido la selección de la gran variedad de maices "modernos" (Zea mays) a partir del teocintle (Zea perennis) en mesoamérica. Pero, también es sabido que la "selección" de nuevos híbridos no ha sido benéfica en todos sus aspectos, como lo han sido la patentación de híbridos transgénicos por Monsanto, el incremento en el uso de elementos tóxicos en la agricultura, la contaminación genética y la pérdida en la biodiversidad de cultivos "nativos" o criollos, el desarrollo de resistencia en insectos y hierbas, efectos nos deseados en otros individuos, entre otros. Todo esto nos lleva a cuestionarnos si es necesario regular y cambiar la forma en que se administra el uso de nuevos híbridos.

Además, es bien sabido que la sociedad en general suele considerar que la palabra "genómica" es un sinónimo de la palabra "transgénica", lo que puede conllevar a un rechazo general de esta metodología por parte de la sociedad. Pero, ¿son los híbridos trangénicos los únicos híbridos involucrados en esta metodología? La respuesta es que no son los únicos híbridos a los que se les puede aplicar esta metodología, ya que de manera convencional se puede obtener nuevos híbridos a partir de la selección de líneas puras (resultantes de la autopolinización), la selección de híbridos mediante métodos de pedrigree (resultantes de polinzación cruzada) y otras propuestas que reesultan ser combinaciones de estos métodos de cruza, como lo son el método a granel ("Bulk Method") y el método de cruzas repetidas ("Back Cross Method"), entre otros, que pueden resultar de cruzar diversas generaciones entre sí.

Finalmente, podemos observar que hay muchas limitantes y otras consideraciones implicadas por el presente proyecto, siendo esto solo "la punta de un iceberg" que puede resultar de gran importancia para la sociedad.

Referencias

[1] Montesinos-López, O.A.; Montesinos-López, A.; Cano-Paez, B.; Hernández-Suárez, C.M.; Santana-Mancilla, P.C.; Crossa, J. 2022. *A Comparison of Three Machine Learning Methods for Multivariate Genomic Prediction Using the Sparse Kernels Method (SKM) Library.* Genes (Basel). 13(12):2279. DOI: 10.3390/genes13122279. PMID: 36553548; PMCID: PMC9778253. https://doi.org/10.3390/genes13081494

[2] Monteverde, E.; Gutierrez, L.; Blanco, P.; Pérez de Vida, F.; Rosas, J.E.; Bonnecarrère, V.; Quero, G.; McCouch, S. 2019. Integrating Molecular Markers and Environmental Covariates To Interpret Genotype by Environment Interaction in Rice (Oryza sativa L.) Grown in Subtropical Areas. *G3 Genes Genomes Genet, 9*, 1519–1531.

[3] Sevilla, S. 2007. Metodología de los estudios de asociación genética. *Insuficiencia cardíaca, 2 (3)*, 111 - 114. ISSN 1852-3862.

[4] Yang, Y.,; Zhu, K.; Xia, H.; Chen, L.; Chen, K. 2014. Comparative proteomic analysis of indica and japonica rice varieties. *Genet Mol Biol.;37(4)*, 652-61. DOI: 10.1590/S1415-47572014005000015. Epub 2014 Oct 21. PMID: 25505840; PMCID: PMC4261965.

[5] Poehlman, J.M. y Sleper, D.A. Capítulo 3. *Breeding Field Crops*, Cuarta Edición, Iowa: Iowa State Press.

[6] VanRaden P.M. 2008. Efficient methods to compute genomic predictions. *J Dairy Sci, 91*, 4414– 4423.