Una comparación entre métodos estadísticos clásicos y técnicas metaheurísticas en el modelamiento estadístico

A contrast study on classic statistical methods and metaheuristic techniques in statistical modeling

Julián A. Acuña Collazos[1], Andrés H. Domínguez Castaño[2], Eliana M. Toro Ocampo[3]
[1]*Facultad de Ciencias, Universidad del Tolima, Ibagué, Colombia*
[2]*Facultad de Ingenierías, Universidad Tecnológica de Pereira, Pereira, Colombia*
[3]*Facultad de Ingeniería Industrial, Universidad Tecnológica de Pereira, Pereira, Colombia*
jualacco@gmail.com
ahdominguez@utp.edu.co
elianam@utp.edu.co

***Resumen*— Este artículo se basa en el problema de selección de variables para representar modelos estadísticos utilizando un algoritmo genético de Chu-Beasley (AGCB). El AGCB utiliza una heurística constructiva en la generación de la población inicial y dos etapas de mejoramiento que funcionan como restricciones para evaluar la calidad del modelo seleccionado. En la literatura especializada se han evidenciado avances con el algoritmo genético tradicional mostrando buenos resultados en el problema de modelamiento estadístico, sin embargo el AGCB aun no ha sido evaluado en la solución de este tipo de problemas. Se muestran cinco métodos de selección de variables en dos grupos. Un grupo consta de tres métodos de selección estadísticos clásicos paso a paso y el otro consta de un algoritmo genético tradicional y un AGCB. Luego, se comparan los resultados obtenidos con base en el ajuste, error estándar y en función de ajuste del modelo seleccionado con dos casos de prueba, donde el algoritmo genético propuesto obtuvo mejor desempeño que las técnicas clasicas.**

***Palabras clave*— Ajuste del modelo, algoritmo genético, criterios estadísticos, error estándar del modelo, métodos de selección, selección de variables.**

***Abstract*— This paper approaches the variables selection problem for representing statistical models using a Chu-Beasley Genetic algorithm (CBGA) based methodology. The CBGA generates the initial population using a constructive heuristic and a two-stage improvement phase that work as restrictions for evaluating the quality of the statistical model selected. In the availiable specialized literature it has been evidenced advances in the traditional genetic algorithm showing good results in the statistical modeling problem. Eventough, the CBGA it has not been yet evaluated for the solution of these type of problems. Five methods are shown for the variable selection split in two groups. The first group consists in three clasic statistical selection step by step methods, and the other group is made up of a traditional genetic algorithm and a CBGA. Then, the fitting, standard error and the selected model fitness function based results are compared using two study cases, where the proposed genetic algorithm has a better performance than the classical techniques.**

***Key Word* —** Model fitting, genetic algorithm, statistical criteria, model standard error, selection methods, variable selection.

I. INTRODUCCIÓN

En el análisis de regresión múltiple, la construcción, evaluación y selección del mejor subconjunto de variables predictoras que expliquen una variable respuesta es un problema importante de la estadística por diversas razones que incluyen [1]:

a. Estimar o predecir a un menor costo al reducir el número de variables sobre las que se recogen datos.

b. Predecir con precisión mediante la eliminación de las variables sin relevancia.

c. Describir un conjunto de datos multivariados con parsimonia. Se dice que un modelo es parsimonioso si consigue ajustar bien los datos pero usando la menor cantidad de variables predictoras posibles.

d. Estimar los coeficientes de regresión con errores estándar pequeños (sobre todo cuando algunas variables predictoras están altamente correlacionadas).

e. Emplear un menor conjunto de variables predictoras de forma que se mitigue el esfuerzo computacional.

Por lo anterior, el estudio de la selección del mejor subconjunto de variables no es un trabajo fácil, especialmente cuando se tiene un gran número de variables predictoras y no se tiene información precisa sobre la relación exacta entre las variables. A veces el número del total de posibles modelos es enorme, (2^k, millones), es decir cuando existen más de k=20 variables predictoras, la evaluación de todas las posibles combinaciones de subconjuntos de variables es una tarea que puede tener un alto costo computacional. Por lo tanto, las técnicas de optimización combinatorial y las estrategias para la selección de

Fecha de Recepción: 26 de Enero de 2012
Fecha de Aceptación: 30 de Abril de 2012

modelos tienen gran importancia y son necesarias para explorar el gran espacio de soluciones [2].

Las estrategias más conocidas y utilizadas para la selección del mejor subconjunto de variables son los métodos "Stepwise", donde el procedimiento se basa en seleccionar el mejor modelo de manera secuencial incluyendo o excluyendo una sola variable predictora en cada paso según criterios de evaluación. Existen tres algoritmos usualmente usados: "*Backward Elimination*" (Eliminación hacia atrás), "*Forward Selection*" (Selección hacia adelante) y "*Stepwise Selección*" (Selección Paso a Paso). Sin embargo estos métodos presentan una argumentación teórica muy escasa en cuanto a la decisión del orden en que las variables participan o se omiten en la conformación del modelo estadístico [3], también en la forma arbitraria en que se asigna la probabilidad a priori para escoger o remover una variable. Por otro, estos métodos solo emplean una búsqueda local, debido a que trabajan en pequeñas áreas del gran espacio de soluciones del problema, por lo tanto, los métodos "*stepwise*" en raras ocasiones encuentran el mejor subconjunto de variables del modelo.

Técnicas metaheurísticas como el Algoritmo Genético [4], son útiles cuando se pretende resolver problemas de optimización para los cuales las técnicas exactas no resultan eficientes o no son aplicables. El algoritmo genético es eficiente en problemas que presentan múltiples óptimos locales [5], facilitando la exploración del espacio de solución en problemas de tamaño considerable en relación con los procedimientos estándar. Esta técnica de optimización, ha sido utilizada en una gran variedad de aplicaciones en campos como la ingeniería, economía, teoría de juegos, ciencias computacionales, mercadeo, biología, medicina, entre otras, logrando ajustarse adecuadamente para cada caso y obteniéndose buenos resultados.

Algunos criterios estadísticos basados en información del modelo como el criterio de información de Akaike (AIC) [6], el criterio de información Bayesiano (BIC) y/o el criterio de información de Schwartz (SIC) [7], que evalúan el grado de calidad de la regresión múltiple según el subconjunto de variables, presentan debilidad para medir la complejidad del modelo a partir del número de variables predictoras en términos de penalidad, la cual es una medida de compensación por el sesgo en la falta de ajuste cuando los estimadores de máxima verosimilitud son utilizados [8]. Sin embargo, no es suficiente medir la complejidad del modelo (término penalidad, por ejemplo 2k en AIC) únicamente con variables predictoras o parámetros del mismo modelo [8]. Es necesario considerar más elementos de juicio para definir y medir la complejidad de la información del modelo seleccionado. En este trabajo se propone utilizar una técnica de optimización combinatorial que sea computacionalmente eficiente en la selección del modelo estadístico con un criterio de evaluación que contenga más propiedades que relacionen e interactúen las componentes de un modelo de regresión.

II. ANÁLISIS DE REGRESIÓN: DIAGNÓSTICOS

Los diagnósticos de regresión se refieren a la clase general de técnicas para la detección de problemas en regresión, en el modelo o en los datos. Estos métodos están diseñados para detectar fallas en los supuestos, observaciones atípicas, deficiencias en el modelo y detección de situaciones en las que las relaciones fuertes entre las variables independientes están afectando los resultados. En este campo de investigación se han hecho algunas publicaciones [9], [10], [11], [12], sin embargo no existe una frontera clara entre la utilidad de estas técnicas con el tiempo. A continuación se mostrarán las técnicas para diagnósticos cuando se presentan problemas de multicolinealidad y para la detección de puntos influenciables.

A. Diagnósticos de colinealidad

El problema de colinealidad en regresión se refiere a que las columnas de la matriz de regresión X pueden estar casi linealmente dependientes o colineales, lo cual conlleva a que $X'X$ esté cerca de ser singular. Entonces la matriz de varianza-covarianza (1) está cerca de la colinealidad, teniéndose un efecto considerable en la precisión. Luego si los coeficientes del modelo de regresión lineal (β) pueden ser estimados y tienen grandes varianzas, las pruebas de estimación de los parámetros del modelo tienen poca influencia y los intervalos de confianza podrían ser muy amplios, haciendo difícil decidir si una variable hace una contribución significativa a la regresión.

$$Var\ (\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad (1)$$

A través del coeficiente de determinación múltiple (R^2) se puede detectar una relación dependiente cuando es cercano a 1 ó 100% para cada par de variables predictoras, sin embargo cuando existen outliers esta medida no es completamente apropiada. Por otro lado, cuando se extiende el caso a más de dos variables predictoras, un conjunto $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k)$ son colineales si para las constantes $(c_0, c_1, \ldots, c_k)$, la siguiente relación: $(c_1 * X_1 + c_2 * X_2 + \ldots + c_k * X_k = c_0)$ se cumple y el R^2_k de una variable de regresión X_k con las demás variables predictoras es cercano a 1, entonces se puede considerar que existe multicolinealidad.

Otra medida para detectar colinealidad es el factor de inflación de la varianza para el k-ésimo coeficiente de regresión (VIF_k). Consideramos el modelo de regresión lineal múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (2)$$

Entonces la varianza del k-ésimo coeficiente de regresión estimado es:

$$Var\ (\hat{\beta}) = \sigma^2 \left(\frac{1}{1 - R_k^2} \right) \left(\frac{1}{S_{X_k X_k}} \right) \qquad (3)$$

La medida $1/(1- R^2_k)$ es denominada el *k*-ésimo factor de inflación de la varianza o (VIF_k) [13]. Si el valor de R^2_k es cercano a 1 entonces la varianza de los parámetros estimados del modelo ($\hat{\beta}_k$) aumenta demasiado. En otras palabras, el *VIF* representa el incremento en la varianza del coeficiente de regresión estimado de una variable predictora debido a la presencia de colinealidad. Una variable predictora con un *VIF* mayor a 10, puede causar colinealidad. Para calcular los *VIF* se utiliza la inversa de la matriz de correlaciones C^{-1} y luego los *VIF's* serán los elementos de la diagonal principal C^{-1}.

B. Influencia estadística

En el análisis de regresión es importante realizar el diagnóstico de influencia estadística para analizar y conocer qué observaciones muestrales afectan en mayor grado el ajuste del modelo de regresión. En la literatura especializada, gran cantidad de autores se han enfocado en medidas de influencia proporcionando metodologías para evaluar el efecto en el ajuste del modelo y/o en el cambio de los coeficientes de regresión estimados al eliminar la *i*-ésima observación del conjunto de datos. Algunas de las más comunes medidas de influencia son: La distancia de Cook D_i [10], el $DFFITS_i$ [9], el $DFBETA_{j(i)}$ [9], el $COVRATIO_i$ [9], la estadística Q_i, entre otros. La estadística Q_i permite evaluar para la i-ésima observación, el cambio en *SCE* cuando el modelo $Y=X\beta+\varepsilon$ se ajusta después de eliminar dicha observación, es decir [12]:

$$Q_i = \frac{\varepsilon_i^2}{(1 - h_{ii})} = SCE - SCE(i) \quad (4)$$

Donde *SCE* es la suma de cuadrados residual cuando el modelo se ajusta con todas las *n* observaciones y $SCE_{(i)}$ es la suma de cuadrados residual cuando el modelo se ajusta sin la *i*-ésima observación.

III. CRITERIOS ESTADÍSTICOS DE SELECCIÓN DE SUBCONJUNTOS DE VARIABLES

Los criterios estadísticos están basados en el principio de parsimonia, donde se recomienda seleccionar un modelo con la suma de los cuadrados residuales pequeños utilizando el mínimo número de variables. No obstante la selección del criterio puede dar lugar a diferentes opciones de tamaño del subconjunto de variables y pueden darse diversos puntos de vista de la magnitud de las diferencias entre los subconjuntos de modelos, siendo esto un aspecto relevante cuando se comparan modelos competentes [14]. Estos criterios pueden ser divididos en tres clases [33]: Criterios de predicción, criterios de información o verosimilitud y criterios de maximización bayesiana con distribución a posteriori de probabilidad.

A. Criterio de información de Akaike: AIC

Akaike en una serie de publicaciones [6], [16], [17], es uno de los pioneros en el campo de la evaluación de modelos estadísticos y aporta a la temática de selección de modelos el criterio de información de Akaike (AIC) definido como:

$$AIC = -2\log L(\hat{\theta}) + 2p \quad (5)$$

Donde $L(\hat{\theta})$ es la función de máxima verosimilitud y *p* es el número de parámetros en el modelo. El criterio precisa que el modelo con el menor valor AIC es seleccionado como el mejor al que se ajustan los datos. La estructura del AIC está compuesta entre la maximización del logaritmo de verosimilitud, es decir ($-2log\, L(\hat{\theta})$), como componente de la falta de ajuste del modelo y *p* como el número de parámetros estimados dentro del modelo como componente de penalidad. La penalidad es una medida de la complejidad o compensación por el sesgo debido a la falta de ajuste cuando los estimadores de máxima verosimilitud son empleados [18]. Para un modelo de regresión lineal múltiple ($Y=X\beta+\varepsilon$), el criterio de información de Akaike (AIC) se define a continuación:

$$AIC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2(k+1) \quad (6)$$

Donde $\hat{\sigma}^2$ es la varianza de los residuales, *k* es el número de variables predictoras en el modelo de regresión y *n* es el número de observaciones de la muestra.

B. Criterio de información de Akaike corregido: AICc

En el criterio AIC definido en la ecuación (1), el sesgo es aproximado por el número de parámetros los cuales son constantes y no tienen variabilidad. Para el modelo de regresión múltiple, la corrección del sesgo del logaritmo de la verosimilitud es calculada como [18]:

$$Sesgo = E_G\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\hat{\theta}) - \int_R \log f(X|\hat{\theta})\,dG(X)\right] = \frac{n(k+1)}{n-k-2} \quad (7)$$

Si se emplea la ecuación (6) para el modelo de regresión múltiple, se puede definir el criterio AICc para una muestra finita, el cual fue propuesto originalmente por [19], como se observa a continuación:

$$AIC_C = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2\frac{n(k+1)}{n-k-2} \quad (8)$$

De manera similar que en el AIC, se selecciona el modelo con el menor valor AICc.

C. Criterio de información bayesiano: BIC

Para mejorar la inconsistencia del criterio AIC, Akaike [17] y Schwarz [7] presentaron un criterio de selección de modelos desde la perspectiva bayesiana. Schwarz estableció que la solución de bayes consiste en seleccionar el modelo con una alta probabilidad a posteriori. Para grandes muestras esta probabilidad a posteriori puede ser aproximada por la expansión de Taylor. Schwarz define el primer término de su criterio como el logaritmo de los estimadores de máxima verosimilitud (MLE's) para el modelo y el segundo término como *p*log(n)*, entonces el criterio de información bayesiano (BIC) es definido como sigue:

$$BIC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + p\log(n) \quad (9)$$

Donde p es el número de parámetros en el modelo y n es el tamaño de muestra. El criterio selecciona el mejor modelo como el que tiene el menor valor BIC.

D. Criterio de complejidad de la información: ICOMP

Bozdogan [2], [8], [18], presenta un nuevo criterio de selección del mejor modelo estadístico basado en la definición de complejidad de un modelo, describiéndola en términos de la interacción entre componentes de un modelo y la información pertinente para su construcción. Luego presenta el enfoque de complejidad de la información ICOMP (*IFIM*) para la evaluación de modelos basado en la complejidad de máxima covarianza. Finalmente para un modelo lineal normal multivariado o no lineal, presenta el criterio de selección de modelos ICOMP (IFIM) definido en forma general como:

$$ICOMP\ (IFIM) = -2\log L(\hat{\theta}) + 2C_1(F^{-1}(\hat{\theta})) \quad (10)$$

Donde, C_1 define la máxima complejidad de la información de F^1 expresada como la matriz inversa de información estimada de Fisher (IFIM) de un modelo, o conocida como la matriz de límite inferior de Crámer-Rao (CRLB). El enfoque de ICOMP (IFIM) aprovecha las propiedades asintóticas óptimas de los estimadores de máxima verosimilitud y utiliza la información basada en la complejidad de la matriz inversa de información de Fisher (IFIM). Finalmente, el criterio ICOMP (IFIM) para un modelo de regresión múltiple, selecciona el mejor modelo como el que tiene el menor valor ICOMP (IFIM) definido de la siguiente forma:

$$ICOMP_{IFIM} = nlog(2\pi) + nlog\left(\widehat{\sigma^2}\right) + n + \cdots$$
$$\ldots (q+1)log\left(\frac{Tr\left(\sigma^2(X'X)^{-1}\right)+\frac{2\sigma^4}{n}}{q+1}\right) - \cdots$$
$$\ldots log\left|\widehat{\sigma^2}(X'X)^{-1}\right| - log\left(\frac{2\widehat{\sigma^4}}{n}\right) \quad (11)$$

IV. MÉTODOS ESTADÍSTICOS DE SELECCIÓN DE SUBCONJUNTOS DE VARIABLES

A. Método de selección backward elimination [21]: SBS

El método inicia con el modelo completo (todas las k variables predictoras). En cada paso se va eliminando una variable del modelo según se cumpla una de las siguientes condiciones:

1) La variable con el menor valor del estadístico F parcial definido como:

$$F_p = \frac{SSR_k - SSR_{k-1}}{MSE_k} \quad (12)$$

Donde SSR_k es la suma de cuadrado de la regresión con k variables, SSR_{k-1} es la suma de cuadrados de la regresión con $k-1$ variables y MSE_k es el cuadrado medio del error del modelo con las k variables. Se calcula el F_p para cada una de las variables que se encuentren en el modelo y se excluye la variable que tiene el F_p más pequeño.

2) La variable que genera la menor reducción en el R^2 al ser descartada del modelo.

3) La variable que tiene el menor coeficiente de correlación parcial en valor absoluto con la variable dependiente.

El proceso del método finaliza cuando se llega a un número prefijado p^* de variables predictoras o cuando el valor del F_p de todas las variables no eliminadas en el modelo es mayor a un valor fijado F_{out} (F_{out} usualmente 4). Es común fijar con anterioridad un nivel de significancia dado α^* (por lo general del 10%) para la prueba "t" o "F" en cada paso y termina el método cuando todos los valores p son menores que α^*. El inconveniente en este método es que una variable que ha sido eliminada del modelo, nunca puede entrar en la regresión de nuevo.

B. Método de selección forward selection [21]: SFS

Este método inicia con un modelo que tiene solo el término constante (ε). Se utiliza la variable predictora con mayor correlación con la variable dependiente en valor absoluto. Si la primera variable no es significativa entonces se tiene el modelo $\hat{Y} = \bar{Y}$ y se detiene el proceso, sino la siguiente variable que entra al modelo cumple con cualquiera de las siguientes condiciones:

1) La variable que tiene el mayor F_p entre las variables que no están incluidas en el modelo.

2) La variable que genera el mayor crecimiento del R^2 al ser incluida en el modelo.

3) La variable que tiene el mayor coeficiente de correlación parcial en valor absoluto con la variable dependiente.

El proceso de este método finaliza cuando se obtiene un número fijado p^* de variables predictoras o cuando el valor del F_p de todas las variables que aún no han sido incluidas en el modelo es menor a un valor fijado F_{in} (F_{in} usualmente igual a 4). Es común fijar con anterioridad un nivel de significancia dado α^* (por lo general del 5%) para la prueba "t" o "F" en cada paso y termina el método cuando todos los valores p de las variables no incluidas son aún mayores que α^*. El problema en este método es que una variable que ha sido incluida en el modelo, nunca puede ser removida de la regresión.

C. Método de selección stepwise selection: SS

Este método propuesto por Efroymson [20] y Draper y Smith [21], combina los métodos SFS y SBS y también es conocido como algoritmo de regresión por pasos (*stepwise regression algorithm*), en el cual comienza con el SFS seguido por el SBS en cada paso. Este algoritmo inicia con el modelo que contiene solo el término constante (ε) y enseguida ejecuta el paso SFS adicionando una sola variable. Luego se aplica el paso SBS el cual remueve una variable si el correspondiente F_p es menor que el F_{out} [22]. Es de notar que en este algoritmo se usan F_{in} y F_{out} con $F_{in} \leq F_{out}$. Esta combinación se repite hasta que ninguna de las variables que no han sido seleccionadas, tengan el grado de importancia necesaria como para ser incluidas en el modelo.

V. MÉTODOS DE SOLUCIÓN: ALGORITMOS GENÉTICOS

El algoritmo genético (AG) pertenece al grupo de las denominadas técnicas metaheurísticas, útiles cuando se pretende resolver problemas de optimización para los cuales las técnicas exactas no resultan eficientes o no son aplicables. Las técnicas metaheurísticas, a diferencia de las heurísticas, poseen mecanismos para escapar de soluciones óptimas locales en su intento por encontrar la solución óptima global. Es de aclarar que ninguna técnica metaheurística garantiza el óptimo global del problema. Comparada con las heurísticas, las metaheurísticas encuentran soluciones muy superiores, con esfuerzos computacionales mayores pero aceptables desde el punto de vista práctico [5]. Los algoritmos genéticos son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos. A lo largo de las generaciones, las poblaciones evolucionan en la naturaleza de acorde con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin en 1859. Por imitación de este proceso, los algoritmos genéticos son capaces de ir creando soluciones para problemas del mundo real. La evolución de dichas soluciones hacia valores óptimos del problema depende en buena medida de una adecuada codificación de los problemas estudiados. Los principios básicos de los algoritmos genéticos fueron establecidos por Holland [4].

A. Algoritmo genético simple: AG

El algoritmo genético simple, también denominado canónico necesita una codificación o representación del problema, que resulte adecuada al mismo. Además se requiere una función de ajuste ó adaptación al problema, la cual asigna un número real a cada posible solución codificada. El algoritmo de ejecución en el trabajo es como sigue: Deben seleccionarse un par de padres para la reproducción, luego dichos padres se cruzan para generar un par de hijos, después este proceso se repite hasta obtener tantos hijos como tamaño de población se tenga definida para el problema, formando una población transitoria sobre la cual actuará un operador llamado mutación para finalmente obtener la nueva generación. En conclusión el AG simple realiza la siguiente secuencia de operaciones:

1. Genera una población inicial, después de elegir el tipo de codificación para representar cada configuración.
2. Calcula la función objetivo de cada configuración de la población y almacena la incumbente (es decir la mejor configuración encontrada durante el proceso).
3. Realiza selección.
4. Realiza recombinación.
5. Realiza mutación y termina de generar la nueva población de la siguiente generación.
6. Si el criterio de parada (o criterios de parada) no se han cumplido el proceso regresa al paso 2.

B. Algoritmo genético Chu-Beasley: AGCB

El algoritmo genético de Chu-Beasley (AGCB) [23], fue diseñado inicialmente para resolver el problema de asignación generalizada, también se tienen reportes de su adaptación a otro tipo de problemas como en sistemas eléctricos con muy buenos resultados. El AGCB está basado en la teoría fundamental de los algoritmos genéticos, pero presenta algunas diferencias que lo hacen muy competitivo para evaluar sistemas de gran tamaño y complejidad. A continuación se mencionan las principales características del AGCB que lo hace un algoritmo más eficiente:

1. Utiliza la función objetivo para identificar el valor de la solución de mejor calidad y maneja la infactibilidad para el proceso de reemplazo de una solución generada a través del proceso de selección-recombinación-mutación por otra que se encuentra en la población actual.
2. A diferencia del AG propuesto por Holland, el algoritmo AGCB sólo genera y sustituye una configuración a la vez en la población, en cada ciclo generacional.
3. Es un algoritmo elitista, ya que un padre será reemplazado por un descendiente en la próxima generación, si y sólo si, el descendiente tiene una función de ajuste de mejor calidad que el padre.
4. Cada configuración que entra a hacer parte de la población debe ser diferente a todos los que conforman la población actual, lo que evita la convergencia prematura a soluciones óptimas locales.
5. Puede incluir una etapa de mejoramiento después de realizar selección, recombinación y mutación. Esto permite explotar la solución descendiente antes de determinar si puede reemplazar a un individuo de la población actual.

C. Implementación del AG al problema de selección de variables

Para un problema de regresión teniendo k variables, el modelo completo puede ser expresado como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \quad (13)$$

.

Donde para el modelo general existen 2^k diferentes posibles subconjuntos para *k* variables. La codificación binaria usada obedece la siguiente regla:

Si la *i*-ésima posición es 0, entonces la i-ésima variable independiente no es incluida en el modelo.

Si la i-ésima posición es 1, entonces la i-ésima variable independiente es incluida en el modelo.

La codificación binaria 1 0 1 1…1 0 1 puede ser representada por el modelo reducido:

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{k-2} X_{k-2} + \beta_k X_k + \varepsilon \quad (14)$$

Siendo la primera posición en la configuración binaria la que representa al intercepto β_0 del modelo de regresión múltiple. Para nuestro problema, considérese un problema de regresión con *k* = 5 variables candidatas más el intercepto.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Supóngase que 2 subconjuntos modelos de regresión fueron escogidos aleatoriamente de la población inicial después de pasar por el proceso de selección (En este trabajo se utiliza la selección tipo ruleta expuesta en [5]) con sus correspondientes AICc como *fitness*:

110001 representa $Y = \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \varepsilon$ con AICc = 38.32.
100101 representa $Y = \beta_0 + \beta_3 X_3 + \beta_5 X_5 + \varepsilon$ con AICc = 34.18.

Luego de aplicar la técnica de ruleta para el operador selección, se debe ejecutar el operador recombinación como se muestra en la Figura 1, correspondiendo a recombinación de un único punto, haciendo el supuesto de que se eligió la posición 3 del cromosoma como punto de recombinación.

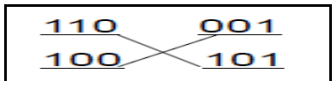

Figura 1. Operador de recombinación de un único punto

La cual resulta en dos nuevas configuraciones:

110101 representa $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_5 X_5 + \varepsilon$ con AICc = 32.25,
100001 representa $Y = \beta_0 + \beta_5 X_5 + \varepsilon$ con AICc = 40.16.

Luego las anteriores configuraciones formaran parte de una población llamada *población transitoria*. El proceso de selección y recombinación se repite hasta obtener tantos individuos como tamaño de población se tenga definido.

Finalmente, el operador mutación debe ser aplicado a la población transitoria para obtener la nueva generación. Este último operador se considera de gran importancia debido a que podría completar alguna variable o variables (según la tasa de mutación) que le hagan falta al proceso de optimización para llegar la solución optima en un determinado momento; en otras palabras este operador genético es el complemento del operador recombinación, ya que puede ayudar a completar el modelo estadístico mejorando el ranking de un determinado individuo ó individuos, aumentando sus posibilidades de producir descendientes de mejor calidad. Además estos individuos pueden tener genes dentro de su configuración que hagan parte del óptimo global aumentando la probabilidad de obtener buenos resultados y en el mejor de los casos el óptimo global.

D. Implementación del AGCB al problema de selección de variables

Para el modelo de regresión teniendo *k* variables (13). La codificación binaria usada en el algoritmo recita:

Si la *i*-ésima posición es 0, entonces la i-ésima variable está presente en el modelo.

Si la *i*-ésima posición es 1, entonces la i-ésima variable está ausente en el modelo.

La generación de la población inicial es de gran importancia en el algoritmo genético. En la literatura especializada se han obtenido buenos resultados cuando se construye una parte de la población inicial de forma heurística y otra parte de manera aleatoria. La generación de la población heurísticamente está basada en la técnica heurística constructiva llamada "*Método ávido aleatorio"*. El método ávido aleatorio [24], es una heurística constructiva basada en una función ávida que guía la entrada de variables en la solución enfocándose en la descomposición de la varianza. Considerando a *x* una variable definida sobre las *n* observaciones, $x'=(x_1, x_2, \ldots, x_n)$, *ng* el número de clases y nn_i el número de casos del grupo *i*, *i=1,….,ng*. Sea:

$\bar{\bar{X}}$: Media de la variable *x* en el conjunto de las *n* observaciones.

$\bar{X}_i$: Media de la variable *x* en los casos de la clase *i*, *i=1,….,ng*.

Cl(j): Clase a la que pertenece el individuo *j*.

Entonces, se definen:

$$VT(x) = \sum_{j=1}^{n} (x_j - \bar{\bar{x}})^2 \quad \text{Variabilidad total} \quad (15)$$

$$VE(x) = \sum_{i=1}^{ng} nn_i (\bar{x}_i - \bar{\bar{x}}) \quad \text{Variabilidad entre grupos.} \quad (16)$$

$$VI(x) = \sum_{j=1}^{n} (x_j - \bar{x}_{cl(j)})^2 \quad \text{Variabilidad intra grupos.} \quad (17)$$

$$F(x) = \frac{VE(x)}{VI(x)} \quad (18)$$

Sea *S* la solución que se va a construir. La función *F* definida en (18) es la que guía el método de selección de variables. No obstante en cada paso no se elije la variable con el mayor valor de *F*. Para esto se construye *L*, denominado *lista de candidatos*, formado por los de mayor valor y aleatoriamente se selecciona uno de la lista *L*. El parámetro α se emplea para controlar el grado de aleatoriedad del método, es decir a mayor valor de α menor aleatoriedad y si α =0, el método es totalmente aleatorio y si $\alpha=1$ el método se vuelve determinístico. La idea de la utilización del método ávido aleatorio, es que se mejore el desempeño con respecto a los métodos de selección determinística. El método ávido aleatorio se describe en [24].

La evaluación de la mejor solución ó la llamada solución incumbente durante el algoritmo genético propuesto la realiza el respectivo criterio de información estadístico que sea utilizado y que fueron mencionados en el apartado III, donde cualquiera de ellos que sea utilizado indica como mejor solución aquella que tenga la menor puntuación. Por otro lado, el método de selección utilizado en este trabajo en el AGCB es el tipo torneo II, es decir, escogiendo *k* individuos en cada uno de los dos torneos, donde en cada torneo compiten entre ellos atreves de su valor de función de adaptación, finalmente se tendrán 2 padres (el ganador de cada torneo) para aplicar el operador recombinación. En el AGCB a diferencia del AG, solo uno de los hijos podrá ser parte de la nueva generación, además la mutación se ejerce solo sobre el descendiente candidato a pertenecer a la siguiente generación. En este trabajo la recombinación se hace en un punto y la tasa de mutación que oscila entre 1% y 5%. Luego, el AGCB tiene una etapa para la mejoría local de un individuo, la cual se aplica antes de la inserción en la población actual del descendiente. Se considera una evaluación y análisis de colinealidad de las variables del modelo que representa el descendiente favorecido, con el fin de identificar variables redundantes que afecten negativamente la estimación del modelo y sus parámetros. En esta evaluación se utiliza el factor de inflación de la varianza (VIF_k) del *k*-ésimo coeficiente de regresión como medida estadística para la detección de colinealidad de variables del modelo. La etapa de reemplazo o inserción en la población está basada en el criterio de diversidad igual a 1, en el que el descendiente debe diferir entre los individuos de la población en al menos 1 gen, además debe ser de mejor calidad que el peor individuo de la población para poder ser parte de la nueva generación [5]. El criterio de parada está definido por un máximo número de generaciones.

Por último, se propone la vinculación de una etapa de mejoramiento externa, en la que posterior al resultado del AGCB se realiza un post-análisis que evalúa las observaciones del modelo que deben ser eliminadas al ser influyentes estadísticamente en la suma de cuadrados de los residuos [25]. Estas dos mejoras, estadísticamente proporcionan un mejor ajuste mediante un modelo reducido tanto en variables como en observaciones que afectan las estimaciones del modelo y sus respectivos parámetros.

VI. RESULTADOS

A continuación se presentan los resultados para dos casos de prueba, en los cuales se evidencia la superioridad de los algoritmos genéticos como métodos de solución con respecto al rendimiento de los métodos estadísticos clásicos paso a paso. También se muestra la eficiencia y mejora significativa del algoritmo genético de Chu-Beasley propuesto comparado con el algoritmo genético simple.

A. Caso 1: Grasa corporal. Para el primer caso se determinó encontrar el mejor subconjunto de variables predictoras de *Y*= Porcentaje de grasa corporal, utilizando *k=13* variables predictoras: X_1=Edad, X_2=Peso, X_3=Altura, X_3=Circunferencia del cuello (cm), X_5=Circunferencia del pecho (cm), X_6=Circunferencia del abdomen (cm), X_7=Circunferencia de la cadera (cm), X_8=Circunferencia del muslo (cm), X_9=Circunferencia de la rodilla(cm), X_{10}=Circunferencia del tobillo (cm), X_{11}=Circunferencia del biceps (extendido) (cm), X_{12}=Circunferencia del antebrazo (cm), X_{13}=Circunferencia de la muñeca (cm). Los datos contienen las estimaciones del porcentaje de grasa corporal determinado por el peso bajo el agua y varias medidas del cuerpo en *n*=252 hombres. La base de datos se encuentra disponible en [26].

Método de selección	R^2	RMSE	No. variables	Variables seleccionadas
Backward Selection	74.4	4.31	7	1,2,4,6,8,12,13
Forward Selection	73.5	4.37	4	2,6,12,13
Stepwise Selection	73.5	4.37	4	2,6,12,13
Algoritmo Genético Simple - AIC	74.5	4.31	8	1,2,4,6,8,10,12,13
Algoritmo Genético Simple - AICc	74.5	4.31	8	1,3,4,6,7,8,12,13
Algoritmo Genético Simple - BIC	73.5	4.37	4	2,6,12,13
Algoritmo Genético Simple - ICOMP	74.1	4.34	7	1,4,6,7,8,12,13
Algoritmo Genético Chu-Beasley - AIC	74.5	4.31	8	1,3,4,6,7,8,12,13
Algoritmo Genético Chu-Beasley - AICc	74.5	4.31	8	1,3,4,6,7,8,12,13
Algoritmo Genético Chu-Beasley - BIC	73.5	4.37	4	2,6,12,13
Algoritmo Genético Chu-Beasley - ICOMP	74.1	4.34	7	1,4,6,7,8,12,13

Tabla 1. Resultados comparativos métodos *Stepwise* vs AG's.

El modelo seleccionado con AG-AIC:

$$Y = -36.6 + 0.0705\ X_1 - 0.129\ X_2 - 0.382\ X_4 + 0.929\ X_6 + 0.226\ X_8 + 0.188\ X_{10} + 0.566\ X_{12} - 1.63\ X_{13} \quad (19)$$

Variable	Coeficiente	Coeficiente de error estándar	T	P	VIF
Constante	-36.578	9.521	-3.84	0,000	
X1: Edad	0.07053	0.03102	2.27	0.024	2.061
X2: Peso	-0.12887	0.03553	-3.63	0,000	14.709
X4: Cuello	-0.3825	0.2239	-1.71	0.089	3.996
X6: Abdomen	0.92866	0.07062	13.15	0,000	7.824
X8: Muslo	0.2262	0.1167	1.94	0.054	5.063
X10: Tobillo	0.1875	0.218	0.86	0.391	1.842
X12: Antebrazo	0.5656	0.1858	3.04	0.003	1.902
X13: Muñeca	-16.346	0.5306	-3.08	0.002	3.311

Tabla 2. Parámetros estimados modelo según AG-AIC.

El modelo seleccionado con AGCB-AIC:

$$Y = 5.00 + 0.0726\ X_1 - 0.141\ X_3 - 0.603\ X_4 + 0.875\ X - 0.351\ X_7 + 0.261\ X_8 + 0.464\ X_{12} - 1.71\ X_{13} \quad (20)$$

Variable	Coeficiente	Coeficiente de error estándar	T	P	VIF
Constante	5.001	7.435	0.67	0.502	
X1: Edad	0.0726	0.03082	2.36	0.019	2.028
X3: Altura	-0.14127	0.08316	-1.7	0.091	1.247
X4: Cuello	-0.6025	0.2151	-2.8	0.005	3.674
X6: Abdomen	0.87536	0.06683	13.1	0.000	6.978
X7: Cadera	-0.3511	0.1191	-2.95	0.004	9.782
X8: Muslo	0.2614	0.1311	1.99	0.047	6.362
X12: Antebrazo	0.4639	0.1859	2.49	0.013	1.897
X13: Muñeca	-17.053	0.4992	-3.42	0.001	2.919

Tabla 3. Parámetros estimados del modelo según AGCB-AIC.

Se puede notar que el modelo seleccionado por el AGCB propuesto, presenta mejora con respecto a la colinealidad entre las variables predictoras, como se observa en la última columna de la tabla 3 (VIF), donde ningún valor del factor de inflación de la varianza de las variables predictoras del modelo, es mayor al umbral del factor de colinealidad igual a 10 utilizado en la parametrización del algoritmo.

Método de selección	Criterio estadístico	Tamaño Población	Mejor criterio estadístico	Generaciones	R^2	RMSE	Número de variables	Variables seleccionadas
Algoritmo genético (AG)	AIC	18	1462.7	61	74.5	4.31	8	1,2,4,6,8,10,12,13
	AICc	12	1462.9	66	74.5	4.31	8	1,3,4,6,7,8,12,13
	BIC	12	1480.5	25	73.5	4.37	4	2,6,12,13
	ICOMP	28	1474.4	43	74.1	4.34	7	1,4,6,7,8,12,13
Algoritmo genético Chu-Beasley (AGCB)	AIC	12	1462.1	19	74.5	4.31	8	1,3,4,6,7,8,12,13
	AICc	10	1462.9	25	74.5	4.31	8	1,3,4,6,7,8,12,13
	BIC	12	1480.5	12	73.5	4.37	4	2,6,12,13
	ICOMP	19	1474.4	20	74.1	4.34	7	1,4,6,7,8,12,13

Tabla 4. Resultados comparativos AG vs AGCB.

Método de selección	R^2	RMSE	Número de variables	Variables seleccionadas	Número de observaciones influyentes	Observaciones por orden de eliminación
AGCB-AIC	80.8	3.58	8	1,3,4,6,7,8,12,13	22	39, 207, 224, 81, 82, 204, 135, 231, 128, 225, 140, 250, 238, 171, 107, 97, 249, 119, 86, 121, 182, 172
AGCB-AICc	80.8	3.58	8	1,3,4,6,7,8,12,13	22	39, 207, 224, 81, 82, 204, 135, 231, 128, 225, 140, 250, 238, 171, 107, 97, 249, 119, 86, 121, 182, 172
AGCB-BIC	80.3	3.58	4	2,6,12,13	25	39, 225, 224, 171, 204, 207, 86, 81, 135, 172, 182, 221, 140, 119, 62, 128, 82, 238, 231, 250, 107, 97, 32, 20, 249
AGCB-ICOMP	81.2	3.63	7	1,4,6,7,8,12,13	20	207, 39, 224, 81, 82, 204, 135, 128, 140, 231, 225, 250, 238, 107, 97, 171, 249, 86, 180, 20

Tabla 5. Resultados de la evaluación de observaciones influyentes

La tabla 5 ilustra los resultados de la evaluación de observaciones influyentes, revelando el AGCB-AIC, AGCB-AICc y AGCB-ICOMP con los mejores ajustes según el R^2, sin embargo el criterio AGCB-ICOMP proporciona un modelo con menor pérdida de grados de libertad dado a que las variables seleccionadas que constituyen el modelo y el número de observaciones influyentes en el ajuste y estimación de parámetros son menores con respecto a los algoritmos que utilizan los criterios AIC y AICc. El mejor modelo seleccionado con AGCB-ICOMP es:

$$Grasa = -4.27 + 0.0819\,Edad - 0.554\,Cuello + 0.896\,Abdomen - 0.335\,Cadera + 0.356\,Muslo + 0.369\,Antebrazo - 2.21\,Muñeca \quad (21)$$

B. Caso 2: Tasas de crecimiento. El segundo caso hace referencia a las mediciones de *Y* =Tasa de crecimiento de *n=72* países durante el período de 1960 a 1980 en función de *k=22* variables predictoras: X_1=Nivel inicial GPD, X_2=Escuela primaria, X_3=expectativa de vida, X_4=Escuela secundaria, X_5=Porcentaje de educación pública, X_6=Rev & coup, X_7=Guerra variable dummy, X_8=Derechos políticos, X_9=Libertad civil, X_{10}=Minería, X_{11}=Organización económica, X_{12}=RFexdist, X_{13}=Inversión en equipos, X_{14}=Inversión en No equipos, X_{15}=Estandarización (BMP), X_{16}=Años de apertura, X_{17}=Edad, X_{18}=Protestantes, X_{19}=Estado de derecho, X_{20}=Crecimiento de la población, X_{21}=Población que trabaja y X_{22}=Fuerza laboral. La base de datos se encuentra disponible en [27].

Método de selección	R^2	RMSE	No. variables	Variables seleccionadas
Backward Selection	80.1	0.89	10	1,3,11,13,14,16,17,18,19,21
Forward Selection	62.9	1.15	5	13,14,16,17,18
Stepwise Selection	62.9	1.15	5	13,14,16,17,18
Algoritmo Genético Simple - AIC	80.5	0.89	13	1,3,6,7,11,13,14,16,17,18,19,20,21
Algoritmo Genético Simple - AICc	80.1	0.89	11	1,3,9,11,13,14,16,17,18,19,21
Algoritmo Genético Simple - BIC	80.1	0.89	11	1,3,10,11,13,14,16,17,18,19,21
Algoritmo Genético Simple - ICOMP	78.2	0.92	10	1,3,10,11,13,14,17,18,19,21
Algoritmo Genético Chu-Beasley – AIC	80.7	0.87	11	1,3,11,12,13,14,16,17,18,19,21
Algoritmo Genético Chu-Beasley – AICc	80.7	0.87	11	1,3,11,12,13,14,16,17,18,19,21
Algoritmo Genético Chu-Beasley – BIC	80.1	0.88	11	1,2,3,11,13,14,16,17,18,19,21
Algoritmo Genético Chu-Beasley – ICOMP	78.5	0.91	10	1,2,3,11,13,14,17,18,19,21

Tabla 6. Resultados comparativos métodos *Stepwise* vs AG's.

Variable	Coeficiente	Coeficiente de error estándar	T	P	VIF
Constante	6.749	1.569	4.30	0,000	
X1	-1.7516	0.2688	-6.52	0,000	5.266
X3	0.08939	0.02204	4.05	0,000	5.932
X11	0.22372	0.09835	2.27	0.027	1.445
X12	-0.004331	0.003108	-1.39	0.169	1.513
X13	17.589	4.813	3.65	0.001	2.569
X14	6.174	2.367	2.61	0.011	1.560
X16	-0.007164	0.003156	-2.27	0.027	1.292
X17	-1.6897	0.5172	-3.27	0.002	2.224
X18	-1.3482	0.5203	-2.59	0.012	1.595
X19	1.3118	0.5291	2.48	0.016	2.930
X21	-2.3477	0.6468	-3.63	0.001	1.396

Tabla 7. Parámetros estimados modelo según AGCB-AIC-AICc.

Método de selección	Criterio estadístico	Tamaño Población	Mejor criterio estadístico	Generaciones	R^2	RMSE	Número de variables	Variables seleccionadas
Algoritmo genético (AG)	AIC	18	202.38	89	80.5	0.89	13	1,3,6,7,11,13,14,16,17,18, 19,20,21
	AICc	18	206.20	49	80.1	0.89	11	1,3,9,11,13,14,16,17,18, 19,21
	BIC	28	225.12	55	80.1	0.89	11	1,3,10,11,13,14,16,17,18, 19,21
	ICOMP	28	219.58	56	78.2	0.92	10	1,3,10,11,13,14,17,18,19,21
Algoritmo genético Chu-Beasley (AGCB)	AIC	10	197.66	79	80.7	0.87	11	1,3,11,12,13,14,16,17,18, 19,21
	AICc	10	203.94	59	80.7	0.87	11	1,3,11,12,13,14,16,17,18, 19,21
	BIC	10	225.02	78	80.1	0.88	11	1,2,3,11,13,14,16,17,18,19, 21
	ICOMP	18	218.44	55	78.5	0.91	10	1,2,3,11,13,14,17,18,19,21

Tabla 8. Resultados comparativos AG vs AGCB.

Método de selección	R^2	RMSE	Número de variables	Variables seleccionadas	Número de observaciones influyentes	Observaciones por orden de eliminación
AGCB-AIC	90.6	0.56	11	1,3,11,12,13,14,16,17, 18,19,21	8	53, 70, 17, 46, 16, 28, 26, 47
AGCB-AICc	90.6	0.56	11	1,3,11,12,13,14,16,17, 18,19,21	8	53, 70, 17, 46, 16, 28, 26, 47
AGCB-BIC	88.5	0.63	11	1,2,3,11,13,14,16,17, 18,19,21	5	53, 70, 17, 46, 16
AGCB-ICOMP	88.5	0.61	10	1,2,3,11,13,14,17, 18,19,21	6	53, 17, 70, 46, 28, 16

Tabla 9. Resultados de evaluación de observaciones influyentes.

Los resultados de la evaluación de observaciones influyentes del AGCB se ilustran en la tabla 9, la cual muestra los criterios AIC y AICc con los mejores ajustes, eliminando del modelo 8 puntos influyentes, según se cumpla con el criterio expuesto en [25]. Sin embargo si el interés es encontrar un modelo parsimonioso que de buenas estimaciones, entonces se cuenta con el modelo seleccionado por el AGCB que tiene como función *fitness* el criterio ICOMP. La ecuación (22) muestra el mejor modelo compuesto por las variables seleccionadas por el AGCB-ICOMP.

$$\begin{aligned} Crecimiento &= 5.47 - 1.79\,Nivel - 0.347\,Primaria \\ &+ 0.08\,Expectativa + 0.262\,Organización + 14.9\,InvEquip \\ &+ 6.94\,InvNoEquip - 1.61\,Edad - 0.945\,Protest \\ &+ 1.60\,EstDerecho - 3.26\,Pobtrab \end{aligned} \quad (22)$$

VII. CONCLUSIONES

Los métodos estadísticos paso a paso por su procedimiento de búsqueda secuencial, seleccionan variables que dan lugar a modelos de regresión con bajo ajuste y a estimaciones de modelos con menor capacidad explicativa de la variabilidad de la variable dependiente y con grandes desviaciones estándar de los errores.

En el algoritmo genético Chu-Beasley, la etapa de mejoramiento interna que evalúa la colinealidad entre las variables predictoras del modelo ayuda a precisar la estimación de los parámetros evitando el incremento de la varianza del coeficiente de regresión estimado de la variable seleccionada cuando hay presencia de colinealidad, a través de una restricción en los factores de inflación de la varianza (*VIF*) de cada variable predictora seleccionada del modelo. Por esta razón el algoritmo presenta un mejor desempeño cuando se presenta este tipo de problemas. En la etapa final del algoritmo genético de Chu-Beasley donde se evalúan las observaciones influyentes, se encuentra una mejora significativa proporcionando soluciones con muy buenos ajustes.

Los resultados obtenidos por el algoritmo genético de Chu-Beasley son de mejor calidad en ajuste y estimación del modelo seleccionado con cada uno de los criterios estadísticos utilizados en relación con los métodos clásicos. Los criterios estadísticos AIC y AICc tienden a seleccionar modelos con mayor número de variables obteniendo mejores ajustes, sin embargo, la ganancia en un mejor ajuste no es suficiente para compensar el factor que penaliza en función de grados de libertad. El criterio ICOMP al considerar más elementos en el factor de penalización, selecciona modelos con menos variables y buenos ajustes y con menor pérdida de grados de libertad. Por otro lado, los resultados generales de la evaluación de observaciones influyentes, los modelos seleccionados por AIC y AICc revelan mayores cantidades de observaciones influyentes, mientras que el criterio ICOMP presenta el menor número de observaciones influyentes a ser eliminadas del modelo para mejorar el ajuste y el error estándar. Los resultados obtenidos muestran que el algoritmo de Chu-Beasley con el criterio de evaluación ICOMP propuesto presenta una alta eficiencia en la mayoría de los casos estudiados.

RECOMENDACIONES

Se sugieren algunas ideas que pudieran extender la presente investigación, como son:

1. Considerar el uso de otros algoritmos combinatoriales.

2. Implementar y/o desarrollar otros tipos de heurísticas constructivas que puedan dar una mayor capacidad de búsqueda al algoritmo.
3. Comparar y aplicar otros criterios de selección de modelos, como por ejemplo métodos bayesianos.
4. Utilizar estas metaheurísticas con otros tipos de modelos de regresión, por ejemplo para modelos de regresión robusta (regresión L1 y regresión M) y modelos lineales generalizados.
5. Analizar el comportamiento del algoritmo genético Chu-Beasley con otras medidas de diagnósticos para problemas de colinealidad y de influencia estadística.

REFERENCIAS

[1]. A.J. Miller. "Selection of subsets of regression variables". *Journal of the Royal Statistical Society*, vol. 147, No. 3, pp. 391, 1984.

[2]. H. Bozdogan. "Statistical data mining and knowledge discovery. Intelligent statistical data mining with information complexity and genetic algorithm". Chapman & Hall/CRC, ISBN 1-58488-344-8, 2004, pp. 44-45.

[3]. D.E. Boyce, A. Farhi, and R. Weischedel. "Optimal subset selection: Multiple regression, interdependence, and optimal network algorithms". New York: Springer-Verlag, ISBN: 9780387069579, 1974, pp. 19.

[4]. J.H. Holland. "Genetic algorithms". *Scientific American*, vol. 267, No.1, pp. 66-72, 1992.

[5]. R. Gallego, A. Escobar, E. Toro. "Técnicas metaheurísticas de optimización". Textos Universitarios, Ed. 2, ISBN: 978-958-722-007-0, 2008, pp. 77-150.

[6]. H. Akaike. "Information theory and an extension of the maximum likelihood principle". Second international symposium on information theory, Académiai Kiadó, Budapest, pp. 267-281, 1973.

[7]. G. Schwarz. "Estimating the dimension of a model". *The Annal of Statistics*, vol. 6, No. 2, pp. 461-464, 1978.

[8]. H. Bozdogan. "Akaike's information criterion and recent developments in information complexity". *Journal of Mathematical Psychology*, ed. 44, pp. 63-76, 2000.

[9]. D.A. Belsley, E. Kuh, R.E. Welsch. "Regression diagnostics". John Wiley & Sons, Inc. ISBN: 0-471-69117-8, 1980, pp. 6-24.

[10]. R.D. Cook. "Detection of influential observation in linear regression". *Technometrics*, vol. 19, No. 1, pp. 15-18, 1977.

[11]. N.R. Draper, J.A. John. "Influential observations and outliers in regression". *Technometrics*, vol. 23, No. 1, pp. 21-26, 1981.

[12]. L.F. Rincón. "Un criterio que compara las estadísticas Q_i y $DF\beta_{j(i)}$ para el análisis de residuales en modelos de rango completo". *Comunicaciones en Estadística*, vol. 2, No. 2, pp. 139-146, 2009.

[13]. D.W. Marquardt. "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation". *Technometrics*, vol. 12, No. 3, pp. 591-612, 1970.

[14]. J.O. Rawlings, S.G. Pantula, D.A. Dickey. "Applied regression analysis". Springer texts in statistics, ed. 2, ISBN: 0-387-98454-2, 1998, pp. 75-78.

[15]. A.J. Miller. "Subset selection in regression". Chapman & Hall/CRC Press Company, Florida, ed. 2, ISBN: 1-58488-171-2, 2002, pp. 111-112.

[16]. H. Akaike. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control*, vol. 19, No. 6, pp. 716-723, 1974.

[17]. H. Akaike. "A Bayesian analysis of the minimum AIC procedure". *The Annal of Statistics*, vol. 30, No. 1, pp. 9-14, 1978.

[18]. H. Bozdogan. "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions". *Psychometrika*, vol. 52, No. 3, pp. 345-379, 1987.

[19]. N. Sugiura, "Further analysis of the data by Akaike's information criterion and the finite corrections". *Communications in Statistics, Theory and Methods*, vol. 7, No. 1, pp. 13-26, 1978.

[20]. M.A. Efroymson. "Multiple regression analysis". *Mathematical Methods for Digital Computers*, vol. 1, pp. 191-203, 1960.

[21]. N.R. Draper, H. Smith. "Applied regression analysis". John Wiley & Sons, Inc, ed. 3, ISBN: 0-471-17082-8, pp. 335-342, 1998.

[22]. G.A.F. Seber, A.J. Lee. "Linear regression analysis". John Wiley & Sons Inc, ed. 2, ISBN: 0-471-41540-5, 2003, pp. 413-418.

[23]. P.C Chu, J. Beasley. "A genetic algorithm for the generalized assignment problem". *Computers & Operation Research*, vol. 24, No. 1, pp. 17-23, 1997.

[24]. O. Gómez, S. Casado, L. Núñez, J. Pacheco. "Resolución del problema de selección de variables cuantitativas mediante GRASP, Aplicación a ratios financieros". XII Jornadas Congreso ASEPUMA, vol. actas 12, No. 1, pp. 5-7, 2004.

[25]. L.F. Rincón. "Un método para determinar un grupo de observaciones influyentes en la *SCE* al ajustar modelos de rango completo". *Comunicaciones en Estadística*, vol. 3, No. 2, pp. 149-162, 2010.

[26]. Doane D.; Mathieson K.; Tracy R.; "Visual Statistics 2.2". Capitulo 17, Datos de grasa corporal, disponible en: http://jhs14.business.msstate.edu/bqa9333/textbook_files/Visual%20Statistics%202.2/Excel%20Files/Databases/Ch%2017%20Multiple%20Regression%20Analysis/BodyFat.xls.

[27]. LeSage J.; Kelley R.; "Spatial Econometrics", Datos de tasas de crecimiento, disponible en: www.spatial-econometrics.com/data/growthley.txt, y www.spatial-econometrics.com/data/growthley.dat.