Predicción del rendimiento de CHC y PBIL aplicados al problema de la selección de la solución deseada

R. Joan-Arinyo
Grup d'Informàtica a l'Enginyeria
E.T.S. d'Enginyeria Industrial de Barcelona
Universitat Politècnica de Catalunya
robert@lsi.upc.edu

M.V. Luzón
Departamento de Lenguajes y Sistemas Informáticos
Escuela Superior de Ingeniería Informática
Universidad de Granada
luzon@ugr.es

E. Yeguas
Departamento de Informática y Análisis Numérico
Escuela Politécnica Superior
Universidad de Córdoba
eyeguas@uco.es

4 de diciembre de 2007

Resumen

El incremento de la complejidad de las instancias del problema de la selección de la solución deseada en resolución de restricciones geométricas supone un aumento considerable del tiempo requerido por las diferentes metaheurísticas para obtener una solución con calidad para el usuario. En tal situación, la predicción del rendimiento de las metaheurísticas puede suponer un importante avance con objeto de acotar tanto el tiempo de ejecución como la calidad de la solución.

La caracterización del rendimiento de las metaheurísticas a partir de distribuciones de longitud de tiempo de ejecución (RLDs) representadas por distribuciones estadísticas continuas conocidas permite parametrizar tal predicción.

En este trabajo se definirán posibles expresiones para predecir el rendimiento óptimo de las metaheurísticas CHC y PBIL ante instancias desconocidas del problema. Para ello se seleccionará un modelo simple de predicción: la regresión lineal simple, partiendo como base de conocimiento del estudio estadístico exhaustivo del comportamiento óptimo de tales algoritmos ante un conjunto suficientemente representativo de instancias correspondientes a un conjunto reducido de tamaños del problema.

1. Introducción

La aplicación de las metaheurísticas CHC, [12], y PBIL, [4], al problema de la selección de la solución deseada en resolución de restricciones geométricas, [5, 20], supone un incremento considerable

del tiempo de cómputo necesario a medida que aumenta el tamaño del problema y la calidad de la solución requerida por el usuario, tal y como se demuestra en [16].

La caracterización del rendimiento de las metaheurísticas a partir de las RLDs, [15], permite comprobar cómo los requerimientos de tiempo se agravan a medida que evoluciona el algoritmo hasta zonas del espacio de búsqueda que permiten obtener mejores soluciones.

Las Figuras 1 y 2 presentan, respectivamente para CHC y PBIL, las RLDs empíricas para tres instancias ejemplo cada una correspondiente a un tamaño distinto del problema: 2^{18}, 2^{19} y 2^{20}. Podemos comprobar como el acercamiento a la solución óptima (*Probabilidad de éxito* $= 1$) en el eje Y, supone para todos los casos un estancamiento y un gran aumento de la longitud de ejecución necesaria en el eje X, medida en número de soluciones evaluadas tal y como se establece en [1], para avanzar hacia una mayor calidad de solución. Así mismo, esta disminución progresiva de la pendiente de la curva con respecto al eje X se hace mayor si la complejidad (tamaño) de la instancia es superior como puede verse en los ejemplos y tal y como se demostró para estos tamaños concretos del problema en [16].

Cada RLD empírica que define la ejecución sobre una instancia del problema de CHC o PBIL, bajo un marco de ejecución óptimo (conjunto de valores estadísticamente óptimo para los parámetros de los algoritmos), puede ajustarse a la función de distribución acumulativa de una distribución estadística continua teórica, [15, 17]. Para la totalidad de instancias de los tamaños del problema 2^{18}, 2^{19} y 2^{20} el rendimiento de los algoritmos queda definido y caracterizado a través de la distribución Gamma, [16].

Toda distribución Gamma se describe a partir de dos parámetros: forma (k) y escala (θ). Podemos obtener a partir de la distribución Gamma estimada para cada algoritmo e instancia concreta del problema (forma: $\hat{k}$ y escala: $\hat{\theta}$) los diferentes descriptores que caracterizan la evolución de este algoritmo sobre dicha instancia. Tales descriptores son, entre otros, los siguientes: media, desviación típica, mediana y percentiles de la longitud de tiempo de ejecución, probabilidades de éxito para tiempos arbitrarios de ejecución y tiempos máximos de ejecución para la obtención de determinadas probabilidades de éxito. Véase [17] para obtener detalles acerca de las expresiones que permiten calcular los mencionados parámetros descriptivos de la evolución de los algoritmos según la distribución Gamma. En [16] se define la influencia de la variación de los valores de forma (k) y escala (θ) en las RLDs y, por tanto, en la evolución del rendimiento de los algoritmos.

Los algoritmos muestran empíricamente un comportamiento similar en su evolución para las diferentes instancias de un mismo tamaño del problema, [16], por lo que podemos aproximar y estimar el rendimiento de un algoritmo sobre las instancias de un mismo tamaño a partir de su RLD Gamma promedio, que llevará asociados sus correspondientes parámetros de forma (k) y escala (θ).

Por otra parte, existe un interés creciente con la complejidad del problema de adaptar la ejecución de los algoritmos en función de los requerimientos del usuario: tiempo de ejecución y calidad de la solución. Es necesaria una estimación del comportamiento de los algoritmos para decidir cuál será su colaboración en la resolución del problema. La parametrización de la evolución y rendimiento de un algoritmo sobre una instancia y, por aproximación, sobre un tamaño del problema a partir de los parámetros de la distribución Gamma supone un punto de partida interesante para la utilización de modelos de predicción simples que permitan caracterizar el rendimiento de los algoritmos sobre instancias desconocidas del problema.

Una de las posibles alternativas viene proporcionada por el modelo de regresión lineal simple. En este trabajo aplicaremos tal modelo para predecir cuáles serán las expresiones analíticas que determinan el rendimiento de los algoritmos CHC y PBIL para cada tamaño del problema, suponiendo que tal rendimiento está determinado por la distribución Gamma. El punto de partida lo constituirán los resultados obtenidos del estudio estadístico exhaustivo de la aplicación óptima de los algoritmos CHC y PBIL a un banco suficientemente representativo de instancias pertenecientes a un conjunto reducido de tamaños del problema: 2^{18}, 2^{19} y 2^{20}, [16]. Puesto que la aplicación

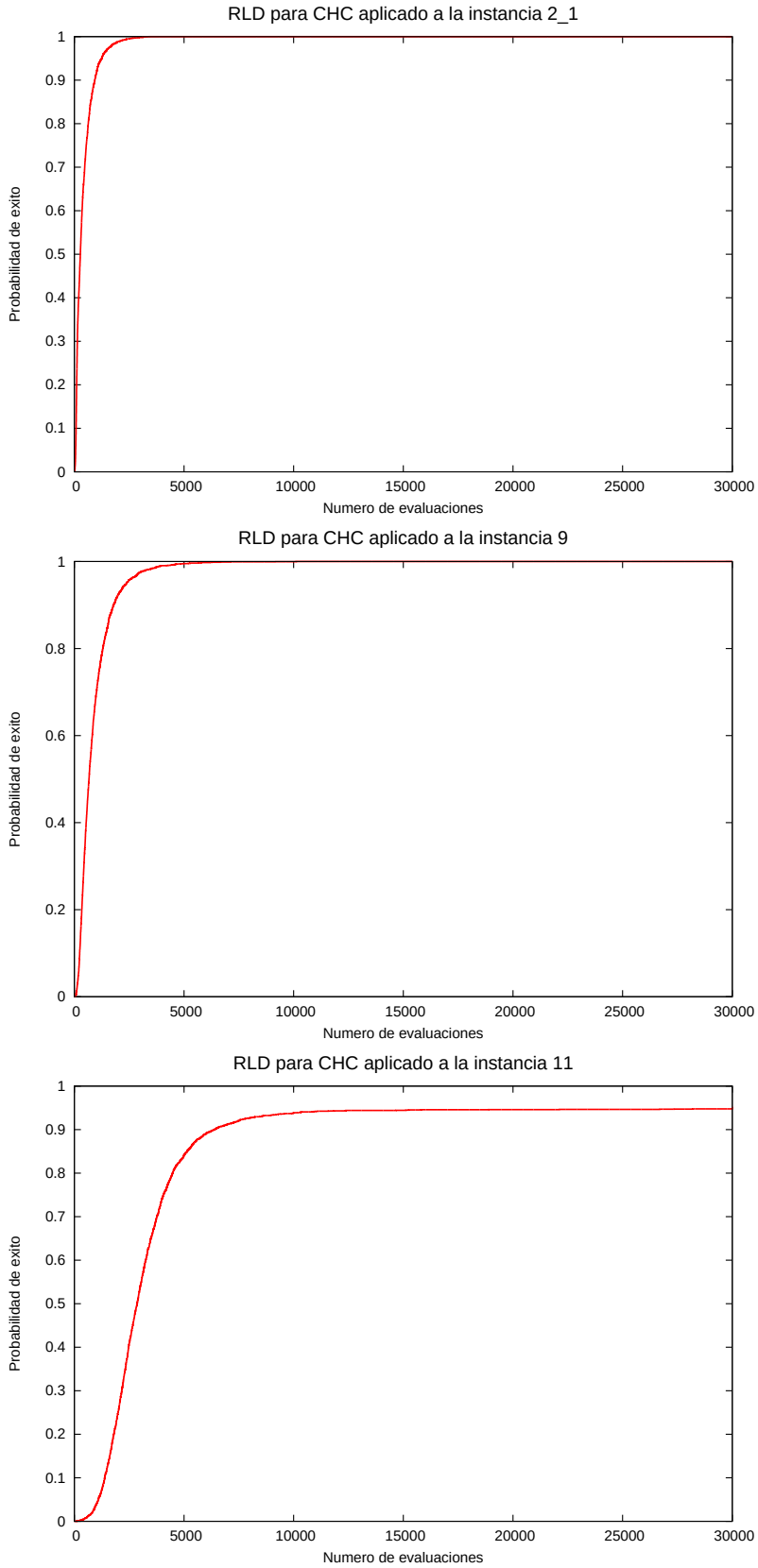

Figura 1: RLDs empíricas correspondientes a la aplicación de CHC a tres instancias ejemplo de tamaños 2^{18} (sup.), 2^{19} (med.) y 2^{20} (inf.).

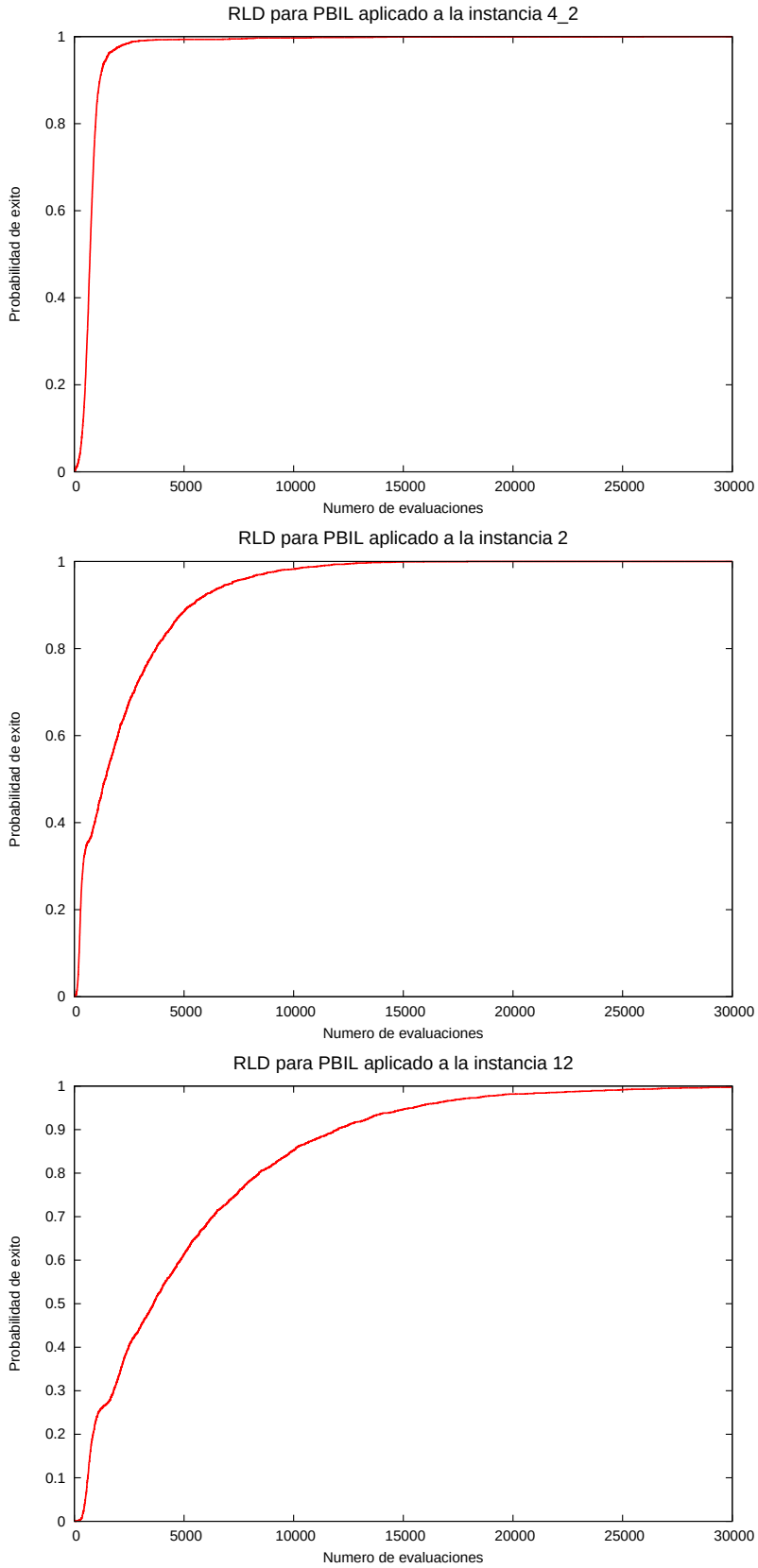

Figura 2: RLDs empíricas correspondientes a la aplicación de PBIL a tres instancias ejemplo de tamaños 2^{18} (sup.), 2^{19} (med.) y 2^{20} (inf.).

de los algoritmos queda definida a partir de un marco de ejecución (individual por instancia y promedio para cada tamaño del problema) definido por un conjunto de parámetros: tamaño de población, umbral diferencia, ratio de divergencia y M mejores individuos para CHC (véase [12]) y tamaño de población, ratio de aprendizaje, probabilidad de mutación y desplazamiento de mutación para PBIL (véase [4]), será necesario deducir también las expresiones analíticas asociadas a los valores de dichos parámetros para realizar la correspondiente generalización de la aplicación de los algoritmos.

A continuación, indicamos cómo se estructura el contenido del presente trabajo. Primero, en la Sección 2 se describen los fundamentos de la regresión lineal simple: definición, hipótesis, contrastes y resultados. La Sección 3 define el diseño procedimental de la aplicación de la regresión lineal simple para la generalización del rendimiento de los algoritmos. Tras ella, en la Sección 4 se presentan los resultados de las diferentes fases de la regresión lineal simple aplicada a los parámetros que definen el modelo de rendimiento Gamma y el marco de ejecución de los algoritmos. La Sección 5 presenta e interpreta las ecuaciones de generalización individual de tales parámetros. Finalmente, las Secciones 6 y 7 establecen respectivamente las conclusiones y trabajos futuros.

2. Regresión lineal simple

Los modelos de regresión, [10, 13], estudian la relación estocástica cuantitativa entre una variable de interés y un conjunto de variables explicativas. Estos modelos son muy utilizados y su estudio conforma un área de investigación clásica dentro de la disciplina de la Estadística desde hace muchos años.

Cuando se estudia la relación entre una variable de interés, variable respuesta o variable dependiente (Y) y un conjunto de variables regresoras (explicativas, independientes) $(X_1, X_2, \ldots, X_n)$, se dice que existe una relación estocástica entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento de éstas permiten predecir con mayor o menor exactitud el valor de la variable respuesta, cuando se cumple la siguiente expresión:

$$Y = m(X_1, X_2, \ldots, X_n) + \varepsilon \tag{1}$$

siendo m la función de regresión desconocida y ε una variable aleatoria de media cero (el error de observación).

El objetivo básico en el estudio de un modelo de regresión es el de estimar la función de regresión, m, y el modelo probabilístico que sigue el error aleatorio ε, ésto es, estimar la función de distribución F_ε de la variable de error. La estimación de ambas funciones se hace a partir del conocimiento de una muestra de las variables en estudio, $\{((X_{1,i}, X_{2,i}, \ldots, X_{n,i}), Y_i) | i = 1, 2, \ldots, M\}$.

Una vez estimadas estas funciones se tiene conocimiento de:

- La relación funcional de la variable respuesta con las variables regresoras, dada por la función de regresión que se define como sigue:

$$m(x_1, x_2, \ldots, x_n) = E(Y/X_1 = x_1, \ldots, Y/X_n = x_n) \tag{2}$$

  Esto permite tener una idea general del comportamiento de la variable respuesta en función de las regresoras.

- Se puede estimar y predecir el valor de la variable respuesta de un individuo del que se conocen los valores de las variables regresoras. Ésto es, de un individuo t se sabe que $X_1 = x_{1,t}, \ldots, X_n = x_{n,t}$, entonces se puede predecir el valor de Y_t y calcular un intervalo de predicción del mismo.

El modelo de regresión más sencillo es el modelo de regresión lineal simple, [11, 23], que estudia la relación lineal entre la variable respuesta (Y) y la variable regresora (X), a partir de una muestra $\{(x_i, Y_i)\}_{i=1}^{M}$, que sigue el siguiente modelo:

$$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \quad \forall i = 1, 2, \ldots, M \tag{3}$$

Por tanto, es un modelo de regresión paramétrico, pues la función de regresión, m, que relaciona a la variable respuesta con las variables regresoras pertenece a una familia paramétrica lineal, y de diseño fijo, puesto que las variables regresoras son valores predeterminados.

2.1. Hipótesis básicas del modelo de regresión lineal simple

Para la aplicación del modelo de regresión lineal simple es necesario que se cumplan las siguientes hipótesis:

- La función de regresión es lineal,

$$m(x_i) = E(Y/x_i) = \alpha_0 + \alpha_1 x_i \quad \forall i = 1, 2, \ldots, M \tag{4}$$

  o, equivalentemente,

$$E(\varepsilon_i) = 0 \quad \forall i = 1, 2, \ldots, M \tag{5}$$

- La varianza es constante (homocedasticidad),

$$Var(Y/x_i) = \sigma^2 \quad \forall i = 1, 2, \ldots, M \tag{6}$$

  o, equivalentemente,

$$Var(\varepsilon_i) = \sigma^2 \quad \forall i = 1, 2, \ldots, M \tag{7}$$

- La distribución es normal,

$$Y/x_i \sim N(\alpha_0 + \alpha_1 x_i, \sigma^2) \quad \forall i = 1, 2, \ldots, M \tag{8}$$

  o, equivalentemente,

$$\varepsilon_i \sim N(0, \sigma^2) \quad \forall i = 1, 2, \ldots, M \tag{9}$$

- Las observaciones Y_i son independientes. Bajo las hipótesis de normalidad, esto equivale a que

$$Cov(Y_i, Y_j) = 0, \quad \text{si } i \neq j \tag{10}$$

  Esta hipótesis en función de los errores sería "los ε_i son independientes", que bajo normalidad, equivale a que

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \text{si } i \neq j \tag{11}$$

2.2. Estimación de la línea de regresión

En el modelo de regresión lineal simple hay tres parámetros que se deben estimar: los coeficientes de la recta de regresión, α_0 y α_1, y la varianza poblacional, σ^2. El cálculo de estimadores para estos parámetros puede hacerse por diferentes métodos, siendo los más utilizados el método de máxima verosimilitud, [2], y el método de mínimos cuadrados, [21]. En nuestro caso, nos centraremos en el segundo de ellos.

El método de mínimos cuadrados tiene como objetivo calcular los estimadores $\hat{\alpha}_0$ y $\hat{\alpha}_1$, a partir de los cuales se pueden calcular las predicciones para las observaciones muestrales, dadas por,

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i \quad \forall i = 1, 2, \ldots, M \tag{12}$$

Definiendo los residuos como la diferencia entre valor observado y valor previsto,

$$e_i = y_i - \hat{y}_i \quad \forall i = 1, 2, \ldots, M \tag{13}$$

los estimadores por mínimos cuadrados se obtienen minimizando la suma de los cuadrados de los residuos, ésto es, minimizando la siguiente función,

$$\Psi(\alpha_0, \alpha_1) = \sum_{i=1}^{M} e_i^2 = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} (y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_i))^2 \tag{14}$$

quedando definidos a partir de las siguientes expresiones:

$$\begin{array}{l} \hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x} \\ \hat{\alpha}_1 = \frac{s_{XY}}{s_x^2} \end{array} \tag{15}$$

siendo $\bar{x}$ e $\bar{y}$ las medias muestrales de X e Y, respectivamente, s_x^2 la varianza muestral de X y s_{XY} la covarianza muestral entre X e Y.

El estimador $\hat{\alpha}_1$, que es la pendiente de la recta de regresión, se denomina coeficiente de regresión y tiene una sencilla interpretación: indica el crecimiento (o decrecimiento) de la variable respuesta Y asociado a un incremento unitario en la variable regresora X.

El estimador $\hat{\alpha}_0$ indica el valor de la ordenada en la recta de regresión estimada para $x = 0$. Presenta menor importancia y en muchos casos no tiene interpretación práctica si el dominio de X es siempre superior a cero.

De las ecuaciones que resultan de la minimización de la función descrita en la ecuación 14 se deduce que los residuos verifican que

$$\begin{cases} \sum_{i=1}^{M} e_i = 0 \\ \sum_{i=1}^{M} e_i x_i = 0 \end{cases} \tag{16}$$

Por tanto, el número de grados de libertad de los residuos es $M - 2$ porque hay M residuos relacionados por dos ecuaciones. Por este motivo, como estimador de σ^2 se utiliza la varianza residual $\hat{s}_R^2$ definida como la suma de residuos al cuadrado dividida por el número de grados de libertad:

$$\hat{s}_R^2 = \frac{1}{M-2} \sum_{i=1}^{M} e_i^2 \tag{17}$$

La desviación típica asociada a $\hat{s}_R^2$, conocida con el nombre de error estándar de estimación $\hat{s}_R$, es una medida de la exactitud de la recta de regresión con respecto a los datos, no estableciendo la desviación típica con respecto a la media de la muestra. Para la cuantificación del error, existe también el denominado error absoluto medio EAM que se define como sigue:

$$EAM = \frac{\sum_{i=1}^{M} |e_i|}{M} \tag{18}$$

2.3. Inferencia en regresión lineal simple

Para determinar si existe una relación significativa entre X e Y, es necesario comprobar el siguiente contraste:

$$C_1 : \begin{Bmatrix} H_0 & \alpha_1 = 0 \\ H_a & \alpha_1 \neq 0 \end{Bmatrix} \tag{19}$$

Si el contraste C_1 no permite rechazar la hipótesis nula H_0 de que la pendiente de la recta de regresión es nula, la recta de regresión sería:

$$Y_i = \alpha_0 + \varepsilon_i \quad \forall i = 1, 2, \ldots, M \tag{20}$$

por lo que no existiría relación lineal entre las variables X e Y.

De la misma forma, se realiza el contraste

$$C_0 : \begin{Bmatrix} H_0 & \alpha_0 = 0 \\ H_a & \alpha_0 \neq 0 \end{Bmatrix} \tag{21}$$

aunque tiene menor interés debido a su escaso significado.

Existen diferentes posibilidades para probar si se cumplen tales hipótesis, aunque la más extendida es el t-test, [19].

Por otra parte, para verificar si el modelo es significativo o no, se hace necesario descomponer la variabilidad de la variable respuesta en variabilidad explicada por el modelo más variabilidad no explicada o residual. Bajo la hipótesis de que existe una relación lineal entre la variable respuesta y la variable regresora, el objetivo es realizar el siguiente contraste de hipótesis:

$$C_F : \begin{Bmatrix} H_0 & E(Y/X = x) = \alpha_0 \\ H_a & E(Y/X = x) = \alpha_0 + \alpha_1 x \end{Bmatrix} \tag{22}$$

Si se acepta H_0 la variable regresora no influye y no hay relación lineal entre ambas variables. En caso contrario, sí existe una dependencia lineal de la variable respuesta respecto de la regresora. Tal verificación se puede realizar mediante ANOVA, [7, 8, 9], donde el estadístico F proporciona una medida global de la bondad de la regresión.

El Contraste de la F (C_F) es un contraste unilateral (de una cola) pero en este modelo proporciona exactamente el mismo resultado que se obtiene por el contraste individual de la t relativo al coeficiente de regresión 1 (C_1), Contraste de la t, mencionado anteriormente.

Una vez ajustada la recta de regresión a la nube de observaciones es importante disponer de una medida que indique la bondad del ajuste realizado y que permita decidir si el ajuste lineal es

suficiente o se deben buscar modelos alternativos. Como medida de bondad del ajuste se utiliza el coeficiente de determinación, definido como sigue

$$R^2 = \frac{scE}{scG} = \frac{\sum_{i=1}^{M}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{M}(y_i - \bar{y})^2} \tag{23}$$

El numerador se denomina Suma de Cuadrados Explicada (scE) y el denominador se suele denotar como Suma de Cuadrados Global (scG). Como $scE < scG$, se verifica que $0 < R^2 < 1$. El coeficiente de determinación mide la proporción de variabilidad total de la variable dependiente respecto a su media que es explicada por el modelo de regresión, por lo que suele expresarse también en %.

Por otra parte, una buena medida de la bondad del ajuste de la recta de regresión viene proporcionada por el coeficiente de correlación lineal muestral (o coeficiente de correlación de Pearson), definido por

$$r = signo(\hat{\alpha}_1)\sqrt{R^2} \tag{24}$$

Por tanto, $r \in [-1, 1]$. Este coeficiente es una buena medida de la bondad del ajuste de la recta de regresión: cuantifica la relación lineal entre las variables X e Y. Evidentemente, existe una estrecha relación entre r y $\hat{\alpha}_1$ aunque estos estimadores proporcionan diferentes interpretaciones del modelo. Es importante estudiar si r es significativo (distinto de cero) ya que ello implica que el modelo de regresión lineal es significativo. Si el valor de $r = +1$ indica una relación lineal exacta positiva (creciente), si por el contrario $r = -1$ la relación sería negativa (decreciente).

2.4. Predicción en regresión lineal simple

Como se comentó anteriormente hay dos objetivos básicos en el ajuste de un modelo de regresión:

- Conocer la relación existente entre la variable respuesta y las variables regresoras. En el caso de la regresión lineal simple se estima la mejor recta de regresión que relaciona la variable Y con la variable X y se cuantifica la importancia de dicha relación por medio del coeficiente de correlación, r.
- Utilizar el modelo de regresión ajustado para predecir el valor de la variable respuesta Y cuando la variable regresora toma un valor determinado, $X = x_t$.

En esta sección se estudia este segundo objetivo. Ésto es, estimada la recta de regresión, ¿cómo predecir el valor de Y sabiendo que la variable regresora toma el valor $X = x_t$? Ante esta pregunta, se deben distinguir dos situaciones diferentes:

- Estimar la media de la distribución condicionada de $Y/X = x_t : E(Y/X = x_t) = m_t$.
- Predecir el valor de la variable respuesta en un individuo de la población en estudio del que se sabe que $X = x_t$. Esto es, predecir un valor de la variable condicionada $Y/X = x_t$.

En nuestro caso, nos centraremos en la primera de las situaciones, pues nuestro objetivo es el de predecir cuál sería el valor promedio de los parámetros que determinan y caracterizan el rendimiento de los algoritmos para un tamaño concreto del problema.

De esta forma, una vez calculada la recta de regresión de la variable Y respecto a X a partir del método de mínimos cuadrados, véase la ecuación 15, se quiere estimar el parámetro $m_t =$

$E(Y/X = x_t)$. Para ello, como estimador se utiliza el que proporciona la recta de regresión, sustituyendo x_t por x en la ecuación de la recta (véase la ecuación 12),

$$\hat{m}_t = \hat{y}_t = \hat{\alpha}_0 + \hat{\alpha}_1 x_t \tag{25}$$

La distribución del estimador $\hat{m}_t$ es normal,

$$\hat{m}_t \sim N\left(\alpha_0 + \alpha_1 x_t, \sigma\sqrt{\frac{1}{M} + \frac{(x_t - \bar{x})^2}{\sum_{i=1}^{M}(x_i - \bar{x})^2}}\right) \tag{26}$$

En la práctica, el estadístico anterior no se puede utilizar para calcular intervalos de confianza de m_t porque σ es desconocido. Por ello, se sustituye σ por su estimador $\hat{s}_R$, deducido de la ecuación 16, y bajo la hipótesis de normalidad se obtiene la siguiente distribución,

$$\frac{\hat{m}_t - m_t}{\hat{s}_R\sqrt{\frac{1}{M} + \frac{(x_t-\bar{x})^2}{\sum_{i=1}^{M}(x_i-\bar{x})^2}}} \sim t_{M-2} \tag{27}$$

La distribución dada permite calcular intervalos de confianza de m_t con un nivel de confianza β, de la siguiente forma,

$$m_t \in \hat{m}_t \pm t_{\beta/2}(M-2)\hat{s}_R\sqrt{\frac{1}{M} + \frac{(x_t - \bar{x})^2}{\sum_{i=1}^{M}(x_i - \bar{x})^2}} \tag{28}$$

2.5. Problemas en el ajuste de un modelo de regresión lineal simple

Al ajustar un modelo de regresión lineal simple se pueden presentar diferentes problemas bien porque no existe una relación lineal entre las variables o porque no se verifican las hipótesis estructurales que se asumen en el ajuste del modelo. Estos problemas son los siguientes:

- Falta de linealidad porque la relación entre las dos variables no es lineal o porque variables explicativas relevantes no han sido incluídas en el modelo.
- Existencia de valores atípicos e influyentes, datos atípicos que se separan de la nube de datos muestrales e influyen en la estimación del modelo.
- Falta de Normalidad debido a que los residuos del modelo no se ajustan a una distribución normal
- Heterocedasticidad, la varianza de los residuos no es constante.
- Dependencia (autocorrelación), existe dependencia entre las observaciones.

Un primer paso para el estudio de estos problemas es la realización de un estudio descriptivo, analítico y gráfico, de la muestra. En particular el gráfico de puntos de la muestra bidimensional, gráfico de dispersión, permite detectar algunos problemas: existencia o no de relación entre las variables X e Y y tipo de la relación, indicios de heterocedasticidad, presencia de puntos atípicos, influencia representativa de una variable regresora no tratada, etc.,

Nombre Modelo	Ecuación Modelo	Transform. X	Transform. Y
Simple	$Y = \alpha_0 + \alpha_1 X$	X	Y
Hiperbólico	$Y = \alpha_0 + \alpha_1 / X$	$1/X$	Y
Inverso	$Y = 1/(\alpha_0 + \alpha_1 X)$	X	$1/Y$
Inverso doble	$Y = 1/(\alpha_0 + \alpha_1 / X)$	$1/X$	$1/Y$
Exponencial	$Y = e^{\alpha_0 + \alpha_1 X}$	X	$\ln Y$
Logarítmico	$Y = \alpha_0 + \alpha_1 \ln X$	$\ln X$	Y
Multiplicativo	$Y = \alpha_0 X^{\alpha_1}$	$\ln X$	$\ln Y$
Raíz X	$Y = \alpha_0 + \alpha_1 \sqrt{X}$	$\sqrt{X}$	Y
Raíz Y	$\sqrt{Y} = \alpha_0 + \alpha_1 X$	X	$\sqrt{Y}$
Curva S	$Y = e^{\alpha_0 + \alpha_1 / X}$	$1/X$	$\ln Y$

Tabla 1: Transformaciones más comunes de las variables X e Y para su ajuste a un modelo lineal.

2.5.1. Transformaciones para la linealidad

La hipótesis básica del modelo de regresión lineal simple es

$$E(Y/X = x) = \alpha_0 + \alpha_1 x, \tag{29}$$

pero en muchos casos en el gráfico de la variable respuesta frente a la variable regresora puede verse que la relación no es de este tipo. A pesar de ello, el modelo de regresión lineal continúa siendo válido en muchas situaciones porque la relación puede convertirse en lineal por medio de una transformación simple en la variable respuesta Y, en la variable regresora X o bien en ambas. Las transformaciones más frecuentes y que han sido contrastadas en este trabajo se presentan en la Tabla 1. En algunos casos, la transformación de las variables del modelo permite resolver problemas como falta de normalidad o heterocedasticidad.

2.5.2. Análisis de residuos

Para comprobar si se verifican las hipótesis estructurales en el ajuste de un modelo lineal, el análisis de residuos juega un papel fundamental. Existen una serie de gráficos sencillos que pueden aportar información relevante sobre los posibles problemas.

El residuo ordinario (e_i) asociado a una observación muestral fue definido en la ecuación 13 como la diferencia entre la observación (y_i) y la predicción ($\hat{y}_i$). Dado que $\sigma^2(e_i)$ no es constante, es difícil identificar las observaciones con residuos grandes. Por ello es usual tipificarlos para obtener los residuos estandarizados (r_i). Los residuos estandarizados tienen media cero y varianza próxima a la unidad, ésto permite distinguir a los residuos grandes.

Una observación con residuo grande se denomina dato atípico (outlier). Cuanto mayor sea r_i más atípica es la observación. Los datos atípicos son de gran importancia porque su inclusión o no en la muestra puede hacer que varíe mucho la recta de regresión estimada. En el modelo de regresión lineal simple es fácil determinar las observaciones que son atípicas y estudiar su influencia en la estimación de la recta ajustada: basta, normalmente, con observar el gráfico de dispersión de la muestra y la recta ajustada.

El análisis gráfico de los residuos estandarizados da una buena idea acerca de si se verifican o no las hipótesis del modelo de regresión. Los gráficos de cajas (Box-Plot), los gráficos P–P y el histograma de los residuos estandarizados proporcionan información sobre la distribución de los mismos. Así mismo, el gráfico de las predicciones frente a los residuos ($\hat{y}_i, e_i$) es el que proporciona una mayor información acerca del cumplimiento de las hipótesis del modelo: linealidad, homocedasticidad y

datos atípicos. Véanse [22] y [14] para comprobar detalles acerca de las diferentes representaciones gráficas.

Por último, en lo que corresponde a la hipótesis de que las observaciones muestrales son independientes, tenemos que, la falta de independencia se produce fundamentalmente cuando se trabaja con variables aleatorias que se observan a lo largo del tiempo. Por ello, una primera medida para tratar de evitar la dependencia de las observaciones consiste en aleatorizar la recogida muestral. Uno de los contrastes válidos y más utilizado para detectarla es el de Durbin-Watson, [3].

El contraste de Durbin-Watson está diseñado para detectar residuos de un modelo de regresión lineal que tienen un coeficiente de autocorrelación de orden uno (ρ) distinto de cero. El contraste es el siguiente

$$C_{DW} : \begin{Bmatrix} H_0 & \rho(\varepsilon_t, \varepsilon_{t+1}) = 0 \\ H_a & \rho(\varepsilon_t, \varepsilon_{t+1}) \neq 0 \end{Bmatrix} \tag{30}$$

El estadístico de Durbin-Watson $\hat{d}$ para este contraste es

$$\hat{d} = \frac{\sum_{t=2}^{M}(e_t - e_{t-1})^2}{\sum_{t=1}^{M} e_t^2} \tag{31}$$

siendo e_t los residuos. Durbin y Watson calcularon la distribución de este estadístico $\hat{d}$, bajo la hipótesis nula y para cada tamaño muestral, M. Para cada nivel de significación β, las tablas Durbin y Watson proporcionan los niveles inferior, d_L, y superior, d_U, de la distribución. Si $d_U < \hat{d} < 4 - d_U$ se acepta H_0, esto es, no hay autocorrelación.

3. Diseño procedimental de la aplicación de la regresión lineal simple para la generalización del rendimiento de los algoritmos

El objetivo de este trabajo es el de generalizar el rendimiento de la aplicación, significativamente óptima desde un punto de vista estadístico, de los algoritmos CHC y PBIL a cualquier instancia del problema de la selección de la solución deseada en resolución de restricciones geométricas. Para ello, partimos de una definición paramétrica de tal rendimiento dado un tamaño del problema: forma (k) y escala (θ) de la distribución Gamma promedio asociada a las RLDs empíricas correspondientes a la ejecución de los algoritmos sobre las diferentes instancias, [16, 15]. Tal definición ha sido obtenida del estudio estadístico exhaustivo de la aplicación de los algoritmos a un conjunto suficientemente representativo de instancias correspondientes a los tamaños del problema: 2^{18}, 2^{19} y 2^{20}.

La aplicación de los algoritmos CHC y PBIL a las diferentes instancias de un tamaño del problema viene determinada por un conjunto de parámetros: tamaño de población, umbral diferencia, ratio de divergencia y M mejores individuos para CHC y tamaño de población, ratio de aprendizaje, probabilidad de mutación y desplazamiento de mutación para PBIL. De ahí que el estudio estadístico exhaustivo del rendimiento de los algoritmos para los bancos de instancias de tamaños 2^{18}, 2^{19} y 2^{20} determinase previamente los valores óptimos de tales parámetros, desde un punto de vista estadísticamente significativo, para cada tamaño del problema. El rendimiento de los algoritmos, caracterizado a través de la distribución Gamma, proviene de un marco de ejecución óptimo, desde un punto de vista estadísticamente significativo.

La generalización de la aplicación de los algoritmos en base al tamaño del problema vendrá definida, por tanto, por un conjunto de expresiones que establecen tanto el modelo de rendimiento (parámetros de la distribución Gamma) como el marco de ejecución (parámetros del algoritmo) de los mismos.

Una de las formas más sencillas de llevar a cabo la generalización comentada vendrá dada por la generalización individual de los parámetros que intervienen tanto en el modelo de rendimiento como en el marco de ejecución. En este trabajo, realizaremos una generalización individual de dichos parámetros utilizando para ello el modelo de regresión lineal simple, teniendo en cuenta su facilidad de aplicación y su significación estadística.

Partiendo de la definición del modelo de regresión lineal simple, Sección 2, para cada algoritmo realizaremos un análisis de regresión por cada parámetro que interviene en el modelo de rendimiento y por cada parámetro que define su marco de ejecución. En cada análisis estudiaremos la relación lineal entre la variable regresora (T), que identificará al tamaño del problema, y la variable respuesta (P), que identificará a cada parámetro concreto. Para cada algoritmo tendremos seis análisis de regresión: cuatro por los parámetros que identifican el marco de ejecución, diferentes para cada algoritmo, y dos por los parámetros del modelo de rendimiento Gamma.

La muestra $\{(t_i, P_i)\}_{i=1}^{NI}$ de partida vendrá constituida por los resultados óptimos, desde un punto de vista estadísticamente significativo, correspondientes a cada parámetro para los tamaños 2^{18}, 2^{19} y 2^{20}, [18, 16]. De cada instancia del problema analizada proviene un elemento, óptimo desde un punto de vista estadísticamente significativo para dicha instancia, de la muestra de cada parámetro. El conjunto de instancias está constituido por $NI = 53$ instancias: 29 de tamaño 2^{18} y 12 de cada uno de los tamaños 2^{19} y 2^{20}.

El proceso de análisis de regresión lineal simple seguido para cada parámetro P correspondiente a cada algoritmo ha sido el siguiente:

- Estudio del cumplimiento previo de las hipótesis básicas de linealidad, homocedasticidad y datos atípicos del modelo de regresión lineal simple a partir de los diagramas de dispersión: T frente a P. Véanse las Secciones 2.1 y 2.5.
- Aplicación de la transformación simple más adecuada (mayor valor de R^2) para la variable respuesta P, para la variable regresora T, o bien en ambas, para garantizar la linealidad en el diagrama de dispersión. Véase la Sección 2.5.1.
- Estimación de los coeficientes $\hat{\alpha}_{0,P}$ y $\hat{\alpha}_{1,P}$ de la línea de regresión a partir del método de mínimos cuadrados. Las predicciones para las observaciones muestrales, en base a la ecuación 12, quedarán dadas por

$$\hat{P}_i = \hat{\alpha}_{0,P} + t_i \hat{\alpha}_{1,P} \quad \forall i = 1, 2, \ldots, NI \tag{32}$$

  Véase la Sección 2.2.
- Análisis de resultados: contrastes y validación esenciales (sin incluir tratamiento de las hipótesis). Véanse las Secciones 2.2 y 2.3.
  - Análisis de la bondad del modelo obtenido: cálculo de los coeficientes de determinación (R^2) y de correlación lineal muestral (r).
  - Análisis de significatividad estadística de los coeficientes $\hat{\alpha}_{0,P}$ y $\hat{\alpha}_{1,P}$:
    - Individual: Contrastes de la t.
    - Conjunta: Contraste de la F.
  - Cuantificación del error: error estándar de estimación $\hat{s}_R$, error absoluto medio EAM.
- Análisis a posteriori de cumplimiento de hipótesis: análisis de residuos. Véase la Sección 2.5.2.
  - Normalidad: histograma de residuos estandarizados y gráfico P–P.
  - Linealidad, homocedasticidad y datos atípicos: gráfico de las predicciones frentes a los residuos estandarizados.

- Independencia de las observaciones muestrales: contraste de Durbin-Watson.

A continuación, mostraremos los resultados de la aplicación del análisis de regresión lineal simple para cada uno de los parámetros.

4. Generalización individual de los parámetros de rendimiento

A continuación presentamos la generalización individual de los parámetros que determinan el rendimiento de los algoritmos CHC y PBIL, modelo Gamma de evolución y marco de ejecución, realizada a partir de un análisis de regresión lineal simple por parámetro y en base al procedimiento descrito en la Sección 3. Comenzamos con un estudio previo del cumplimiento de las hipótesis básicas a partir de los diagramas de dispersión.

4.1. Estudio previo del cumplimiento de las hipótesis: diagramas de dispersión y transformaciones

La Figura 3 muestra, como ejemplo, los diagramas de dispersión correspondientes a los parámetros forma (k) del modelo de rendimiento Gamma de CHC y escala (θ) del modelo de rendimiento Gamma para el algoritmo PBIL.

Podemos observar como en el diagrama de dispersión correspondiente al parámetro k se percibe claramente la relación lineal entre el tamaño (eje X) y el parámetro forma (eje Y). Además, se aprecia la existencia de homocedasticidad, pues existe una amplitud similar en el eje Y del conjunto de puntos constituido por todos los valores de k correspondientes a un mismo tamaño T. Finalmente, no se aprecian valores atípicos en el diagrama.

En lo que se refiere al diagrama de dispersión correspondiente al parámetro de escala para PBIL, se aprecia una cierta desviación de la linealidad, así como una disminución de la amplitud de los valores de tal escala a medida que se incrementa el tamaño del problema T. No se advierte, por otra parte, la presencia de valores atípicos.

El fenómeno ocurrido para el parámetro de escala para el algoritmo PBIL se ha repetido para otros parámetros de ambos algoritmos, por lo que ha sido necesario recurrir a las transformaciones, de la variable regresora (tamaño del problema) o de las variables dependientes (parámetros), para garantizar la linealidad y homocedasticidad. Para la generalización individual de cada parámetro P se ha seleccionado la transformación que maximiza el coeficiente de determinación (R^2).

Las Tablas 2 y 3 establecen, respectivamente para CHC y PBIL, el modelo lineal asociado a la transformación seleccionada, de las presentadas en la Tabla 1, junto al coeficiente de determinación obtenido para la generalización individual de cada parámetro. Puede comprobarse como todo modelo lineal resultado de las correspondientes trasformaciones queda respaldado por un elevado valor de coeficiente de determinación: para una mayoría de los parámetros para ambos algoritmos la variabilidad de la variable dependiente viene explicada en un valor superior al 80 % por la variable regresora y para el resto es, en todo caso, muy superior al 60 %. La Figura 4 presenta el diagrama de dispersión correspondiente a la transformación del parámetro de escala de PBIL para el que ahora sí se cumplen los supuestos de linealidad y homocedasticidad, algo que no ocurría originariamente como pudo comprobarse en la Figura 3.

Se verifican mediante los diagramas de dispersión los supuestos previos del modelo de regresión lineal: linealidad, homocedasticidad e inexistencia de valores atípicos (el supuesto de normalidad será verificado a posteriori), para los diferentes parámetros de PBIL y CHC, tanto relativos al marco de ejecución como al modelo de rendimiento.

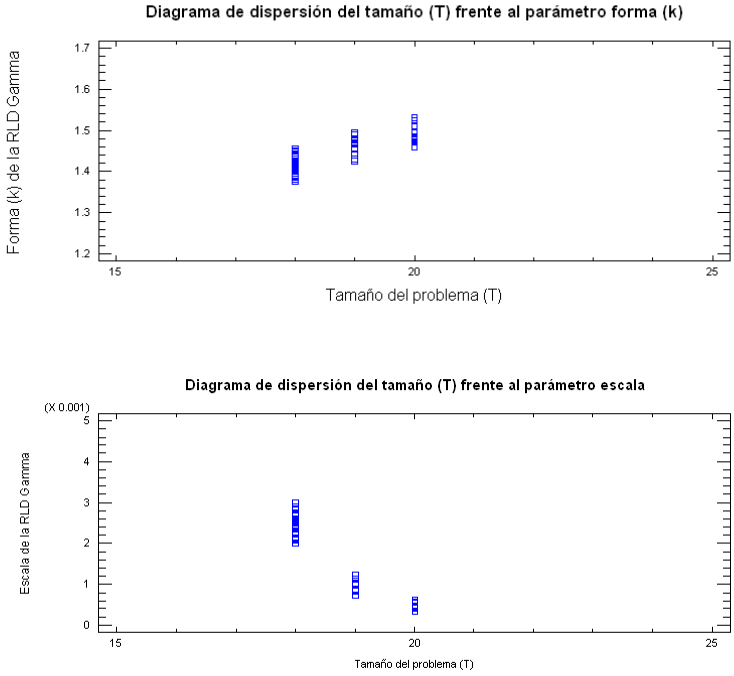

Figura 3: Diagramas de dispersión correspondientes a los parámetros del modelo de rendimiento Gamma: forma para CHC (sup.) y escala para PBIL (inf.).

CHC		
Parámetro	**Modelo lineal**	R^2 **(%)**
Marco de ejecución		
Tamaño de población	Simple	93.8806
Umbral diferencia	Simple	84.6110
Ratio de divergencia	Logarítmico	64.5952
M mejores individuos	Simple	82.9036
Modelo de rendimiento Gamma		
Forma	Simple	70.2094
Escala	Exponencial	91.5930

Tabla 2: Modelo lineal seleccionado para cada parámetro que determina el rendimiento del algoritmo CHC en su relación con el tamaño del problema.

PBIL		
Parámetro	**Modelo lineal**	R^2 (%)
Marco de ejecución		
Tamaño de población	Simple	80.9857
Ratio de aprendizaje	Exponencial	72.3596
Probabilidad de mutación	Raíz P	97.9928
Desplazamiento de mutación	Raíz P	97.8747
Modelo de rendimiento Gamma		
Forma	Simple	82.8621
Escala	Exponencial	95.2714

Tabla 3: Modelo lineal seleccionado para cada parámetro que determina el rendimiento del algoritmo PBIL en su relación con el tamaño del problema.

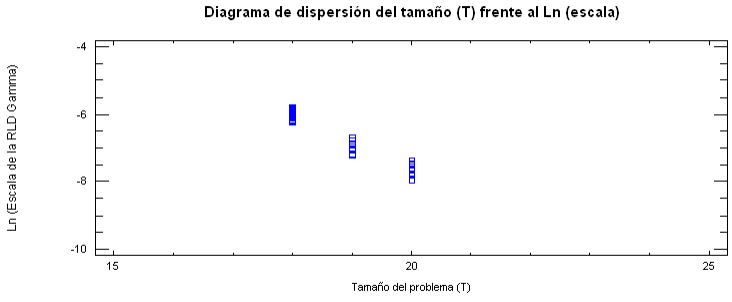

Figura 4: Diagrama de dispersión correspondiente a la transformación de la variable dependiente (parámetro) escala para PBIL.

CHC			
Parámetro (P)	**Ecuación modelo lineal**	$\hat{\alpha}_{0,P}$	$\hat{\alpha}_{1,P}$
Marco de ejecución			
Tamaño de población (PS)	$PS = \hat{\alpha}_{0,PS} + \hat{\alpha}_{1,PS}T$	19.780700	0.380564
Umbral diferencia (D)	$D = \hat{\alpha}_{0,D} + \hat{\alpha}_{1,D}T$	0.841461	0.211591
Ratio de divergencia (DR)	$DR = \hat{\alpha}_{0,DR} + \hat{\alpha}_{1,DR}\ln T$	0.135549	0.062776
M mejores individuos (M)	$M = \hat{\alpha}_{0,M} + \hat{\alpha}_{1,M}T$	1.360900	0.013641
Modelo de rendimiento Gamma			
Forma (k)	$k = \hat{\alpha}_{0,k} + \hat{\alpha}_{1,k}T$	0.645961	0.042728
Escala (θ)	$\theta = e^{\hat{\alpha}_{0,\theta}+\hat{\alpha}_{1,\theta}T}$	9.444350	-0.829201

Tabla 4: Estimación de coeficientes de la línea de regresión para los parámetros que intervienen en el rendimiento del algoritmo CHC.

PBIL			
Parámetro (P)	**Ecuación modelo lineal**	$\hat{\alpha}_{0,P}$	$\hat{\alpha}_{1,P}$
Marco de ejecución			
Tamaño de población (PS)	$PS = \hat{\alpha}_{0,PS} + \hat{\alpha}_{1,PS}T$	7.739150	1.650500
Ratio de aprendizaje (LR))	$LR = e^{\hat{\alpha}_{0,LR}+\hat{\alpha}_{1,LR}T}$	-1.633900	-0.002972
Probabilidad de mutación (MP)	$MP = \hat{\alpha}_{0,MP} + \hat{\alpha}_{1,MP}\sqrt{T}$	0.011871	0.003236
Desplazamiento de mutación (MS)	$MS = \hat{\alpha}_{0,MS} + \hat{\alpha}_{1,MS}\sqrt{T}$	0.017716	0.001375
Modelo de rendimiento Gamma			
Forma (k)	$k = \hat{\alpha}_{0,k} + \hat{\alpha}_{1,k}T$	0.190940	0.069186
Escala (θ)	$\theta = e^{\hat{\alpha}_{0,\theta}+\hat{\alpha}_{1,\theta}T}$	8.907130	-0.829866

Tabla 5: Estimación de coeficientes de la línea de regresión para los parámetros que intervienen en el rendimiento del algoritmo PBIL.

4.2. Estimación de los coeficientes de la línea de regresión y validación del modelo

A continuación, pasamos a la estimación de los coeficientes de la línea de regresión en base al modelo lineal asociado a cada parámetro. Para ello, se ha utilizado el método de mínimos cuadrados. Las Tablas 4 y 5 presentan, respectivamente para CHC y PBIL, los resultados de tal estimación, teniendo en cuenta los modelos lineales correspondientes a los diferentes parámetros, véanse las Tablas 2 y 3.

Tras la estimación de coeficientes, proseguimos por la validación y análisis de los resultados. En primer lugar, analizaremos la bondad del modelo obtenido a partir de los coeficientes de determinación (R^2) y de correlación lineal muestral (r). Las Tablas 6 y 7 muestran, respectivamente para CHC y PBIL, los valores de R^2 y r para los diferentes parámetros que intervienen en el rendimiento de los algoritmos. Para todos los parámetros de ambos algoritmos, tales valores son significativos. La significatividad de R^2 determinó la elección del modelo lineal para cada parámetro como ya se indicó anteriormente. Por otra parte, la cercanía de r a 1 o -1 hace patente la presencia de relaciones lineales positivas o negativas desde un punto de vista estadísticamente significativo. El signo de r es coherente con el tipo de parámetro y sus posibilidades de evolución en función del tamaño del problema: véanse [12] para el marco de ejecución de CHC, [4] para el marco de ejecución de PBIL y [16] para el modelo de rendimiento Gamma en relación con los algoritmos

CHC		
Parámetro	R^2 (%)	r
Marco de ejecución		
Tamaño de población	93.8806	0.9689
Umbral diferencia	84.6110	0.9198
Ratio de divergencia	64.5952	0.8037
M mejores individuos	82.9036	0.9105
Modelo de rendimiento Gamma		
Forma	70.2094	0.8379
Escala	91.5930	-0.9570

Tabla 6: Resultados del análisis de la bondad del modelo para el algoritmo CHC.

PBIL		
Parámetro	R^2 (%)	r
Marco de ejecución		
Tamaño de población	80.9857	0.8999
Ratio de aprendizaje	72.3596	-0.8506
Probabilidad de mutación	97.9928	0.9899
Desplazamiento de mutación	97.8747	0.9893
Modelo de rendimiento Gamma		
Forma	82.8621	0.9103
Escala	95.2714	-0.9761

Tabla 7: Resultados del análisis de la bondad del modelo para el algoritmo PBIL.

CHC y PBIL.

Dentro de la validación y análisis de resultados, proseguimos, en segundo lugar, con el análisis de la significatividad estadística de los coeficientes que se estimaron y presentaron en las Tablas 4 y 5. El proceso se realizará individualmente para cada coeficiente, contrastes de la t, y de forma conjunta a partir del contraste de la F.

Las Tablas 8 y 9 muestran como ejemplo, respectivamente para CHC y PBIL, el resultado de los contrastes de la t, [10], correspondientes a uno de los parámetros que determinan el rendimiento de los algoritmos. Para cada contraste se muestra el parámetro de la recta correspondiente, la estimación del valor del mismo, el error estándar y el estadístico T junto al p-valor que determina el rechazo o no de la hipótesis nula de que el parámetro de la recta se anula. En todos los casos, el p-valor es inferior a 0,05, por lo que garantizamos con un nivel de significación del 95 % el rechazo de la hipótesis nula, y por tanto, existe una relación significativa entre la variable regresora (tamaño del problema) y la variable dependiente (tamaño de población para CHC y ratio de aprendizaje para PBIL). Ello se ha verificado para los contrastes correspondientes a los diferentes parámetros del marco de ejecución y el modelo de rendimiento Gamma de CHC Y PBIL.

En lo que corresponde al análisis de la significatividad estadística conjunta de los coeficientes, las Tablas 10 y 11 presentan, respectivamente para CHC y PBIL, un ejemplo del contraste de la F (Tabla ANOVA, [7]) aplicado a los modelos lineales correspondientes a los parámetros tamaño de población del algoritmo CHC y ratio de aprendizaje del algoritmo PBIL. Puesto que el p-valor correspondiente al estadístico F es en todo caso inferior a 0,05, podemos rechazar la hipótesis nula H_0 con un nivel de significación del 95 % y concluir que existe una dependencia lineal de la variable respuesta respecto de la regresora. Los resultados obtenidos aquí se han generalizado para los diferentes parámetros de los dos algoritmos: marco de ejecución y modelo de rendimiento

CHC: Tamaño de población (TP)					
Contraste	Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
C_0	$\hat{\alpha}_{0,TP}$	19.780700	1.885470	10.491100	0.000000
C_1	$\hat{\alpha}_{1,TP}$	0.380564	0.013605	27.971700	0.000000

Tabla 8: Resultados del análisis de la significatividad individual de los coeficientes del modelo lineal correspondiente al parámetro tamaño de población del algoritmo CHC.

PBIL: Ratio de aprendizaje (LR)					
Contraste	Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
C_0	$\hat{\alpha}_{0,LR}$	-1.633900	0.035655	-45.825400	0.000000
C_1	$\hat{\alpha}_{1,LR}$	-0.002972	0.000257	-11.554800	0.000000

Tabla 9: Resultados del análisis de la significatividad individual de los coeficientes del modelo lineal correspondiente al parámetro ratio de aprendizaje del algoritmo PBIL.

Gamma.

Por último, para finalizar la validación de los diferentes modelos lineales, sin tener en cuenta la verificación del cumplimiento de hipótesis a posteriori, es necesario comprobar la cuantificación del error: error estándar de estimación y error absoluto medio. Las Tablas 12 y 13 muestran, respectivamente para CHC y PBIL, el error estándar de estimación ($\hat{s}_R$) y el error absoluto medio (EAM) para cada parámetro que determina el rendimiento de los algoritmos: marco de ejecución y modelo de rendimiento. Se puede observar como los valores de desviación estándar estimada ($\hat{s}_R$) y error absoluto medio (EAM) son coherentes con el intervalo de valores asociado a cada parámetro y con la progresión estimada en base a los modelos lineales correspondientes, véanse las Tablas 4 y 5. Así mismo, la evolución de los parámetros mantendría lo indicado en [12] para CHC, en [4] para PBIL y en [16] para el modelo de rendimiento Gamma.

4.3. Análisis de residuos

Como conclusión del análisis de regresión lineal, el análisis de residuos permitirá contrastar las hipótesis del modelo a posteriori.

En lo que respecta a la hipótesis de normalidad, aunque ésta deberá cumplirse teniendo en cuenta el teorema del límite central, [7], puesto que disponemos de una muestra suficientemente significativa para el análisis de regresión, hemos realizado los correspondientes gráficos de residuos. La Figura

CHC: Tamaño de población (TP)					
Análisis de varianza: Tabla ANOVA					
Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	P-Valor
Modelo	136985	1	136985	782.41	0.00
Residuo	8929.08	51	175.08		
Total (Corregido)	145914	52			

Tabla 10: Resultados del análisis de la significatividad conjunta de los coeficientes del modelo lineal correspondiente al parámetro tamaño de población del algoritmo CHC.

PBIL: Ratio de aprendizaje (LR)					
Análisis de varianza: Tabla ANOVA					
Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	P-Valor
Modelo	8.3487	1	8.3487	133.5100	0.0000
Residuo	3.1891	51	0.0625		
Total (Corregido)	11.5378	52			

Tabla 11: Resultados del análisis de la significatividad conjunta de los coeficientes del modelo lineal correspondiente al parámetro ratio de aprendizaje del algoritmo PBIL.

CHC		
Parámetro	$\hat{s}_R$	EAM
Marco de ejecución		
Tamaño de población	13.2318	10.4467
Umbral diferencia	1.0207	0.8337
Ratio de divergencia	0.0702	0.0551
M mejores individuos	0.8431	0.7431
Modelo de rendimiento Gamma		
Forma	0.0232	0.0201
Escala	0.0020	0.0012

Tabla 12: Resultados del análisis de la cuantificación del error en el modelo lineal para el algoritmo CHC.

PBIL		
Parámetro	$\hat{s}_R$	EAM
Marco de ejecución		
Tamaño de población	12.3270	10.7598
Ratio de aprendizaje	0.0050	0.0046
Probabilidad de mutación	0.0190	0.0120
Desplazamiento de mutación	0.0197	0.0144
Modelo de rendimiento Gamma		
Forma	0.0262	0.0213
Escala	0.0015	0.0012

Tabla 13: Resultados del análisis de la cuantificación del error en el modelo lineal para el algoritmo PBIL.

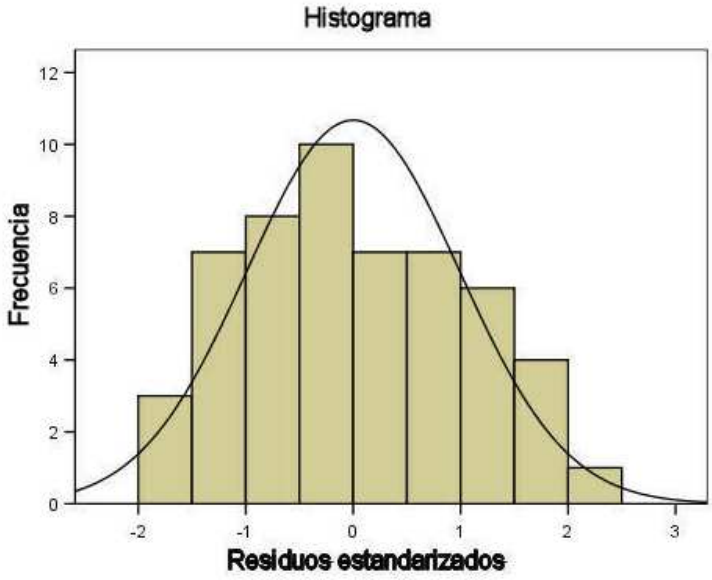

Figura 5: Histograma de residuos estandarizados correspondiente a la variable dependiente (parámetro) forma para PBIL.

5 presenta el histograma de residuos estandarizados y la Figura 6 presenta el gráfico P–P Normal de residuos estandarizados para el modelo lineal del parámetro forma del algoritmo PBIL. En la primera de las Figuras se puede observar la adecuación del histograma a la distribución normal y en la segunda de ellas el ajuste a una línea recta de los puntos correspondientes a la probabilidad acumulada observada frente a la probabilidad acumulada esperada de la distribución normal. Teniendo en cuenta lo establecido en [22, 14], podemos indicar a través de los gráficos que el modelo lineal correspondiente al parámetro forma de PBIL cumple la hipótesis de normalidad. Los resultados han sido similares para el resto de parámetros de rendimiento de ambos algoritmos.

Por otra parte, el cumplimiento a posteriori de las hipótesis de linealidad, homocedasticidad y datos atípicos será verificado a partir de la elaboración de los gráficos de las predicciones frente a los residuos estandarizados. La Figura 7 presenta el gráfico correspondiente para el parámetro forma del modelo de rendimiento del algoritmo PBIL. En el mismo no se detectan puntos atípicos ni ningún tipo de desviación ni descompensación de los puntos que invite a pensar en falta de linealidad o heterodasticidad. Esto es lo que ha ocurrido para el resto de parámetros de ambos algoritmos. Tal y como se establece en [22, 14] las nubes de puntos son adecuadas si existe una distribución homogénea de las mismas, sin presencia de formas, dispersión de puntos ni descompensaciones de los diferentes conjuntos de puntos correspondientes a las diferentes predicciones.

Finalmente, la verificación de la hipótesis de la independencia de las observaciones muestrales permitirá concluir que todas las hipótesis del modelo de regresión lineal se cumplen a posteriori. El contraste de Durbin-Watson ha sido utilizado para esta comprobación. Las Tablas 14 y 15 muestran, respectivamente para CHC y PBIL, los resultados del contraste de Durbin-Watson (estadístico de Durbin-Watson $\hat{d}$ y p-valor asociado) para los modelos lineales correspondientes a los diferentes parámetros que intervienen en el rendimiento de CHC y PBIL. Teniendo en cuenta lo establecido en [3], dado que en todo caso el p-valor es superior a 0.05, no podemos rechazar la hipótesis nula con un 95 % de probabilidad y, por tanto, no existen indicios de autocorrelación serial en los residuos desde un punto de vista estadísticamente significativo.

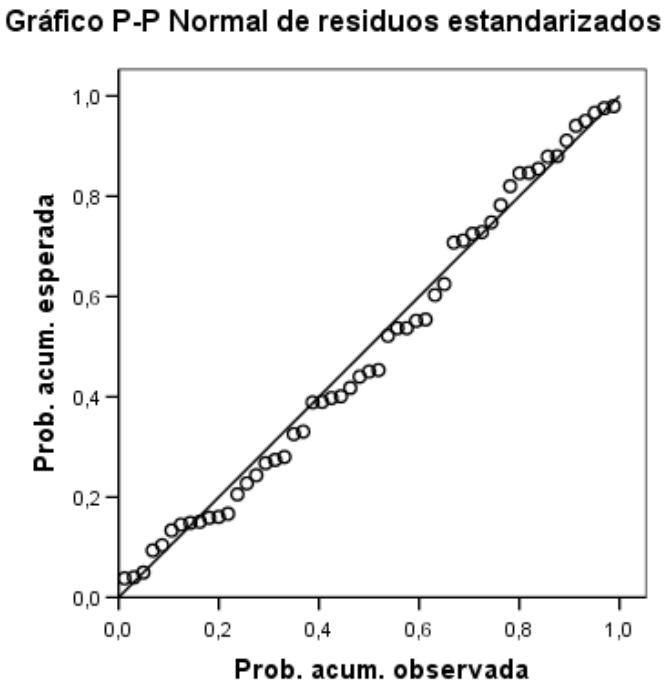

Figura 6: Gráfico P–P Normal de residuos estandarizados correspondiente a la variable dependiente (parámetro) forma para PBIL.

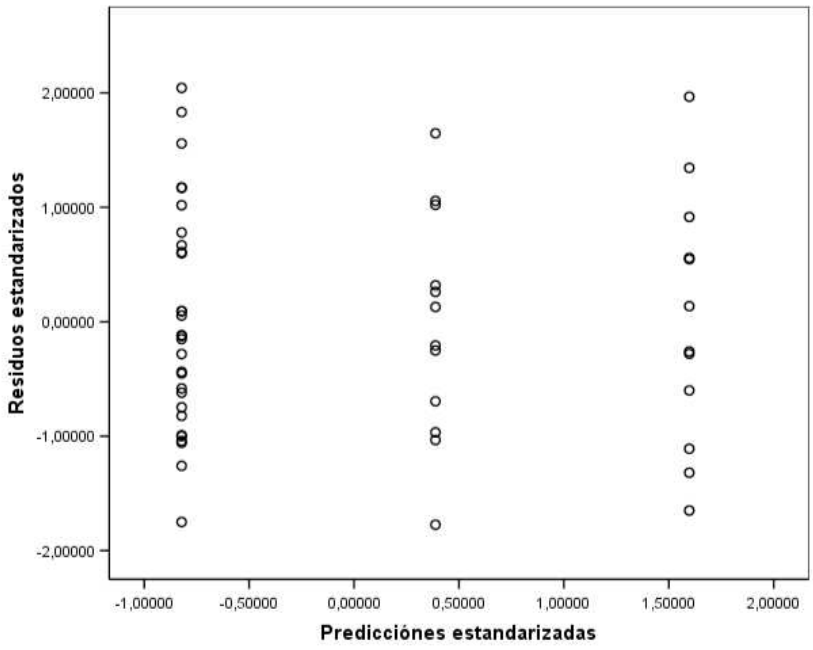

Figura 7: Gráfico de predicciones frente a residuos estandarizados correspondiente a la variable dependiente (parámetro) forma para PBIL.

CHC		
Parámetro	$\hat{d}$	**P-Valor**
Marco de ejecución		
Tamaño de población	1.87634	0.3272
Umbral diferencia	1.82711	0.2404
Ratio de divergencia	2.10111	0.3297
M mejores individuos	1.66662	0.0977
Modelo de rendimiento Gamma		
Forma	1.67377	0.0893
Escala	1.72235	0.1223

Tabla 14: Resultados del contraste de Durbin-Watson para los modelos lineales asociados a los diferentes parámetros de rendimiento del algoritmo CHC.

PBIL		
Parámetro	$\hat{d}$	**P-Valor**
Marco de ejecución		
Tamaño de población	1.63345	0.0806
Ratio de aprendizaje	1.80593	0.2177
Probabilidad de mutación	1.55517	0.0481
Desplazamiento de mutación	2.38075	0.0706
Modelo de rendimiento Gamma		
Forma	1.83719	0.2300
Escala	1.66854	0.0862

Tabla 15: Resultados del contraste de Durbin-Watson para los modelos lineales asociados a los diferentes parámetros de rendimiento del algoritmo PBIL.

5. Interpretación de los resultados de la generalización

Una vez realizado el análisis de regresión y demostrada la validez del ajuste del modelo para los diferentes parámetros, pasaremos a presentar e interpretar los resultados de la generalización con objeto de utilizarlos en una posterior aplicación de los algoritmos CHC y PBIL sobre instancias desconocidas del problema de la selección de la solución deseada. Distinguiremos entre CHC y PBIL.

5.1. CHC

La Figura 8 presenta el ajuste de la muestra de cada parámetro que constituye el marco de ejecución de CHC a su modelo lineal correspondiente establecido en la Tabla 4. Para cada parámetro la curva de regresión está rodeada por dos curvas muy cercanas que delimitan un intervalo de confianza del 95 % para la misma. A su vez, el conjunto de puntos que constituyen la muestra de cada parámetro está delimitado por dos curvas que definen un intervalo de predicción del 95 % (el 95 % de los puntos se ubican en el interior de dicho intervalo). En virtud de la definición de cada uno de los parámetros que constituyen el marco de ejecución de CHC, [12], y teniendo en cuenta la estimación de coeficientes en la regresión lineal de la Tabla 4, la expresión correspondiente a la generalización individual de cada uno de ellos se muestra en la Tabla 16 y se describe a continuación:

- *Tamaño de la población (PS)*: es el número de cromosomas que componen la población que evoluciona en la búsqueda. Se trata, por tanto, de un número entero positivo y de ahí que tomemos la parte entera inferior en la ecuación de generalización. El rango de valores sobre el que se mueve no está definido, no es finito. Tal y como se establece en la bibliografía y lo indica la ecuación de generalización, la población siempre es un número entero positivo y se incrementa a medida que aumenta la complejidad del problema. El incremento para este parámetro entre tamaños del problema próximos es muy reducido, lo que mantiene la coherencia con las características del mismo.

- *Umbral diferencia (D)*: referencia para indicar el máximo grado de parecido permisible entre dos cromosomas para el cruce. Sus valores son números enteros positivos, de ahí que tomemos la parte entera inferior en la ecuación de generalización. Sólo recibe valor en su inicialización, pues en el discurrir del algoritmo, su valor evoluciona tras la reinicialización en función de una expresión. El valor definido como más adecuado en la literatura es de $T/4$. La ecuación de generalización establece que el umbral diferencia siempre es un número entero positivo y se incrementa a medida que aumenta la complejidad del problema. Los valores que recibe en función del tamaño no presentan grandes diferencias con los indicados en la bibliografía.

- *Ratio de divergencia (DR)*: valor que indica el porcentaje de bits, en tanto por 1, que han de ser alterados del mejor cromosoma actual de la población para formar cada uno de los cromosomas del resto de la población en la reinicialización. El valor que puede adquirir es real y corresponde al intervalo $[0, 1]$, de ahí que hayamos establecido el límite en 1 a partir de la función mínimo, pues un valor mayor del mismo supone, por parte del algoritmo, la misma interpretación que éste. La ecuación de generalización se mantiene en el dominio definido para el parámetro y establece un incremento del mismo insignificante a medida que aumenta el tamaño del problema coherente con la definición proporcionada en la bibliografía.

- *M individuos mejores (M)*: valor que indica en una de las posibilidades de reinicialización, el número de cromosomas que permanecerán en la población en la reinicialización, los M mejores. El resto serán generados aleatoriamente. El valor que puede adquirir es un número entero entre 1 y el número de cromosomas total albergado en la población. En la literatura no hay definido un valor concreto como más adecuado, solamente se establece que este valor no ha de albergar una gran cantidad de cromosomas. A esto precisamente es a lo que obedece

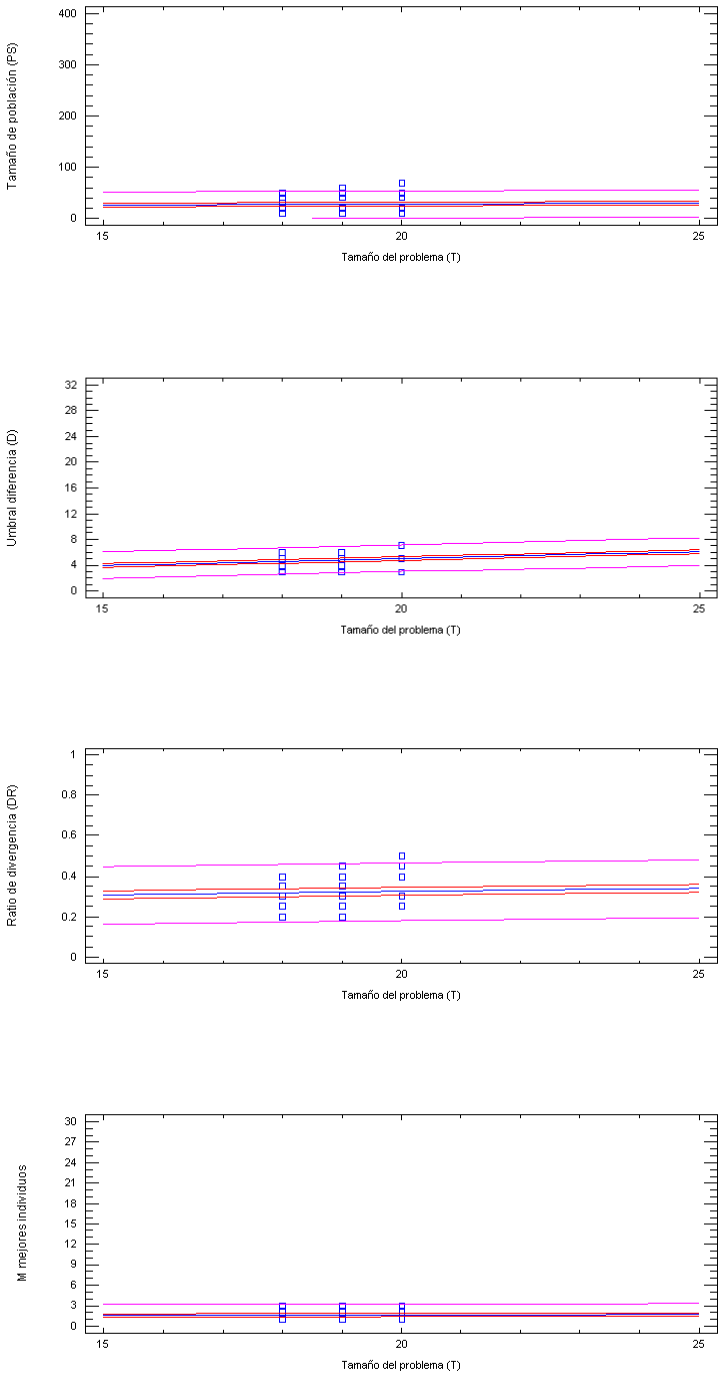

Figura 8: Modelo ajustado para los parámetros que constituyen el marco de ejecución de CHC: tamaño de población (PS), umbral diferencia (D), ratio de divergencia (DR) y M mejores individuos (M).

CHC: *Marco de ejecución*	
Parámetro (P)	**Ecuación generalización**
Tamaño de población (PS)	$\widehat{PS} = \lfloor 19{,}780700 + 0{,}380564T \rfloor;\ \forall T > 0;\ T \in N$
Umbral diferencia (D)	$\hat{D} = \lfloor 0{,}841461 + 0{,}211591T \rfloor;\ \forall T > 0;\ T \in N$
Ratio de divergencia (DR)	$\widehat{DR} = \text{mín}(0{,}135549 + 0{,}062776 \ln T, 1);\ \forall T > 0;\ T \in N$
M mejores individuos (M)	$\hat{M} = \lfloor 1{,}360900 + 0{,}013641T \rfloor;\ \forall T > 0;\ T \in N$

Tabla 16: Generalización individual en función del tamaño del problema de los parámetros que definen el marco de ejecución del algoritmo CHC.

CHC: *Modelo de rendimiento Gamma*	
Parámetro (P)	**Ecuación generalización**
Forma (k)	$\hat{k} = 0{,}645961 + 0{,}042728T;\ \forall T > 0;\ T \in N$
Escala (θ)	$\theta = e^{9{,}444350 - 0{,}829201T};\ \forall T > 0;\ T \in N$

Tabla 17: Generalización individual en función del tamaño del problema de los parámetros que definen el modelo de rendimiento Gamma del algoritmo CHC.

la ecuación de generalización con una pendiente de muy reducida magnitud cuyo resultado es siempre un entero positivo y en torno a los valores recomendados por la bibliografía.

Por otra parte, la Figura 9 presenta, de la misma forma que para los parámetros anteriores, el ajuste de la muestra de cada parámetro que constituye el modelo Gamma de rendimiento de CHC a su modelo lineal correspondiente establecido en la Tabla 4. Teniendo en cuenta la definición de cada uno de los parámetros que constituyen el modelo de rendimiento Gamma de CHC, [16, 6], y la estimación de coeficientes en la regresión lineal de la Tabla 4, la expresión correspondiente a la generalización individual de cada uno de ellos se muestra en la Tabla 17. Ambos parámetros son números reales positivos, lo que se mantiene en las ecuaciones de generalización. Para el problema concreto que nos ocupa, los valores de los mismos evolucionan coherentemente con el análisis previo de los mismos realizado en [16]: mayores dificultades de evolución del algoritmo y hallazgo de soluciones válidas a medida que aumenta la complejidad del problema de la selección de la solución deseada, con el parámetro de forma (k) mostrando un incremento progresivo reducido mientras que los cambios en el parámetro de escala (θ) son mucho más notorios en su decremento.

Partiendo de las ecuaciones de generalización de los parámetros que determinan el marco de ejecución y el modelo de rendimiento de CHC, la predicción de valores para los mismos se realizará mediante la sustitución del tamaño correspondiente a la instancia que se desee resolver. Para conocer detalles acerca de la definición de intervalos de confianza de predicción véase la Sección 2.4.

5.2. PBIL

La Figura 10 presenta, de la misma forma que se mostró para CHC, el ajuste de la muestra de cada parámetro del marco de ejecución de PBIL a su modelo lineal correspondiente establecido en la Tabla 5. Teniendo en cuenta la definición de cada uno de los parámetros que constituyen el marco de ejecución de PBIL, [4], y la estimación de coeficientes en la regresión lineal de la Tabla 5, la expresión correspondiente a la generalización individual de cada uno de ellos se muestra en la Tabla 18 y se describe a continuación:

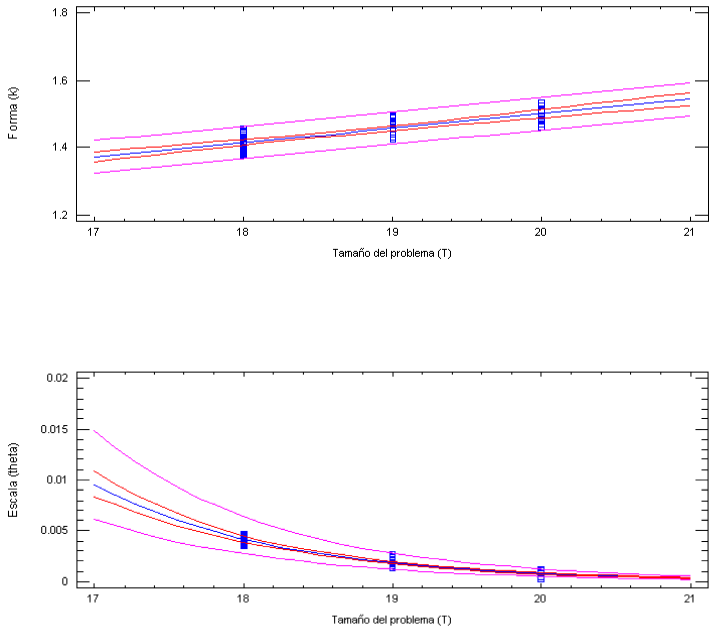

Figura 9: Modelo ajustado para los parámetros que constituyen el modelo de rendimiento Gamma de CHC: forma (k) y escala (θ).

PBIL:*Marco de ejecución*	
Parámetro (P)	**Ecuación generalización**
Tamaño de población (PS)	$\widehat{PS} = \lfloor 7{,}739150 + 1{,}650500T \rfloor$; $\forall T > 0$; $T \in N$
Ratio de aprendizaje (LR))	$\widehat{LR} = e^{-1{,}633900 - 0{,}002972T}$; $\forall T > 0$; $T \in N$
Probabilidad de mutación (MP)	$\widehat{MP} = \text{mín}(0{,}011871 + 0{,}003236\sqrt{T}, 1)$; $\forall T > 0$; $T \in N$
Desplazamiento de mutación (MS)	$\widehat{MS} = \text{mín}(0{,}017716 + 0{,}001375\sqrt{T}, 1)$; $\forall T > 0$; $T \in N$

Tabla 18: Generalización individual en función del tamaño del problema de los parámetros que definen el marco de ejecución del algoritmo PBIL.

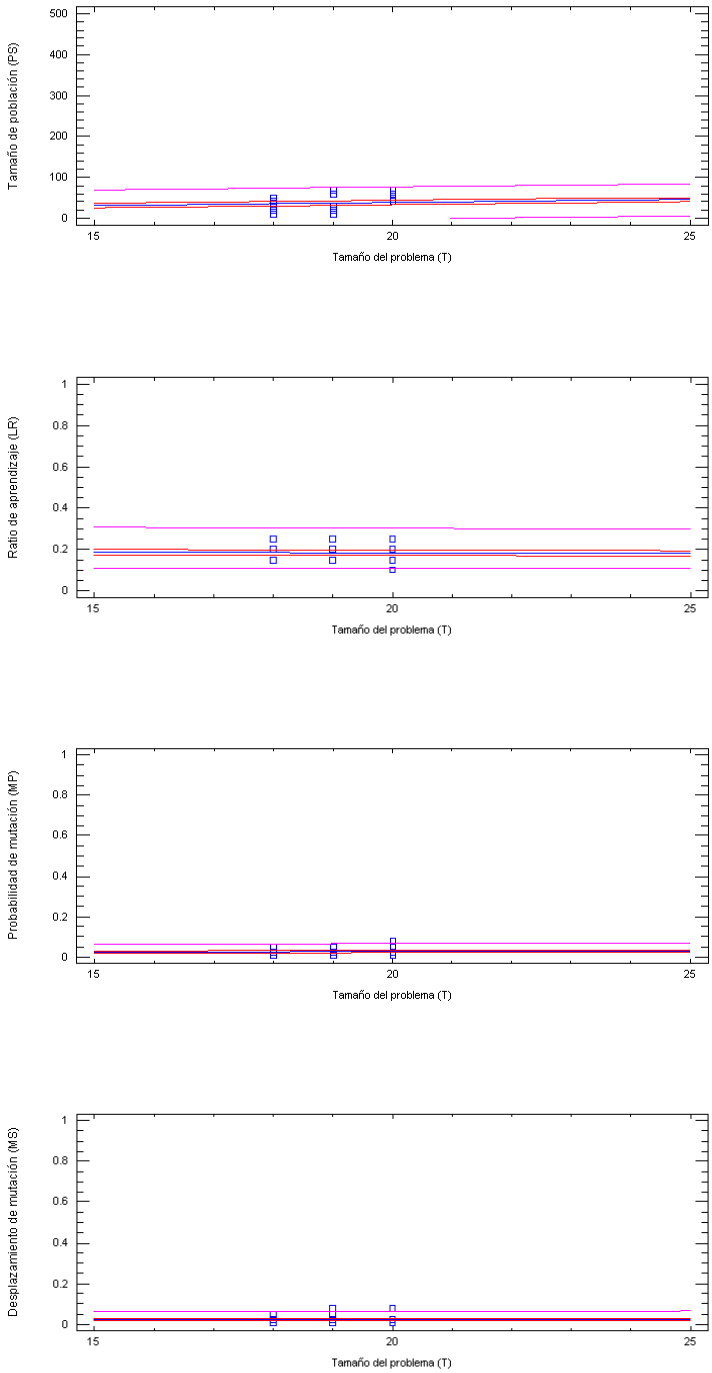

Figura 10: Modelo ajustado para los parámetros que constituyen el marco de ejecución de PBIL: tamaño de población (PS), ratio de aprendizaje (LR), probabilidad de mutación (MP) y desplazamiento de mutación (MS).

- *Tamaño de población (PS)*: indica el número de muestras, soluciones o cromosomas a generar en cada iteración del algoritmo. Hace el papel del tamaño de la población en el algoritmo PBIL, en el que, como ya sabemos, carecemos de una población físicamente presente en la evolución del algoritmo, pues ésta está representada por el vector de probabilidades. El rango de valores sobre el que se mueve no está definido, no es finito. La definición de la ecuación de generalización es, por tanto, similar a la del tamaño de población en CHC. Con dicha definición, mantenemos el valor de este parámetro como un entero positivo y, así mismo, un incremento en el valor del mismo con el aumento de la complejidad del problema, tal y como se describe en la bibliografía. Tal incremento es superior al del tamaño de población de CHC, algo coherente con los experimentos anteriores establecidos en [16].

- *Ratio de aprendizaje (LR)*: se utiliza en la regla de actualización del vector de probabilidades. Con su valor regula la rapidez con que el vector prototipo de probabilidades se acerca a la mejor solución encontrada. Tiene una función similar a la que presenta en el aprendizaje competitivo. El rango de valores sobre el que se mueve es el intervalo $[0, 1]$, aunque los valores preferibles como puede apreciarse en la bibliografía son pequeños para permitir una evolución escalonada hacia la mejor solución y evitar la caída en óptimos locales. La ecuación de generalización obtenida mantiene la presencia de valores cada vez más reducidos de este parámetro ante el incremento de complejidad del problema, respetando su dominio.

- *Probabilidad de mutación (MP)*: indica la probabilidad con la que habrá de ser mutado cada gen del vector de probabilidades. Permite introducir la diversificación en el proceso de evolución del algoritmo. Los valores que toma corresponden al intervalo $[0, 1]$, aunque los valores preferibles, como puede deducirse de lo establecido en la bibliografía, son valores que permitan insertar cada cierto tiempo la diversificación en la búsqueda que permita la evolución ante óptimos locales. A medida que la complejidad del problema se incrementa, la necesidad de diversificación es mayor, por lo que se ha de incrementar progresivamente tal y como lo establece la ecuación de generalización. Para disponer como límite del parámetro al valor 1 hemos insertado la función mínimo, aunque un valor mayor proporcionaría el mismo efecto en el algoritmo.

- *Desplazamiento de mutación (MS)*: indica el valor de desplazamiento del gen del vector de probabilidades una vez que se ha decidido mutar. Este desplazamiento podrá ser en un sentido u otro aleatoriamente decidido. Su valor también corresponde al intervalo $[0, 1]$, pero los valores muy altos para tamaños pequeños del problema son prohibitivos al enfocar la búsqueda en demasía hacia una región concreta del espacio de búsqueda tal y como se puede observar en la bibliografía. Con la ecuación de generalización definida, en la que la pendiente de la recta es de magnitud muy reducida y aparece el término tamaño del problema (T) bajo la raíz cuadrada, aseguramos un crecimiento insignificante del parámetro a medida que se produce un pequeño aumento del tamaño del problema. La introducción de la función mínimo garantiza que mantenemos como valor máximo del parámetro 1, aunque un valor mayor proporcionaría el mismo efecto en el algoritmo.

Por otra parte, la Figura 11 presenta, de la misma forma que para CHC, el ajuste de la muestra de cada parámetro que constituye el modelo Gamma de rendimiento de PBIL a su modelo lineal correspondiente establecido en la Tabla 5. Teniendo en cuenta la definición de cada uno de los parámetros que constituyen el modelo de rendimiento Gamma de PBIL, [16, 6], y la estimación de coeficientes en la regresión lineal de la Tabla 5, la expresión correspondiente a la generalización individual de cada uno de ellos se muestra en la Tabla 19. La interpretación es similar a la de los parámetros del modelo de rendimiento Gamma de CHC. Respeta la relación entre los modelos de rendimiento de ambos algoritmos, analizada en [16], por la que PBIL tiene menor capacidad de evolución ante el estancamiento a medida que aumenta la complejidad del problema.

Tomando como punto de partida las ecuaciones de generalización de los parámetros que determinan el marco de ejecución y el modelo de rendimiento de PBIL, la predicción de valores para los

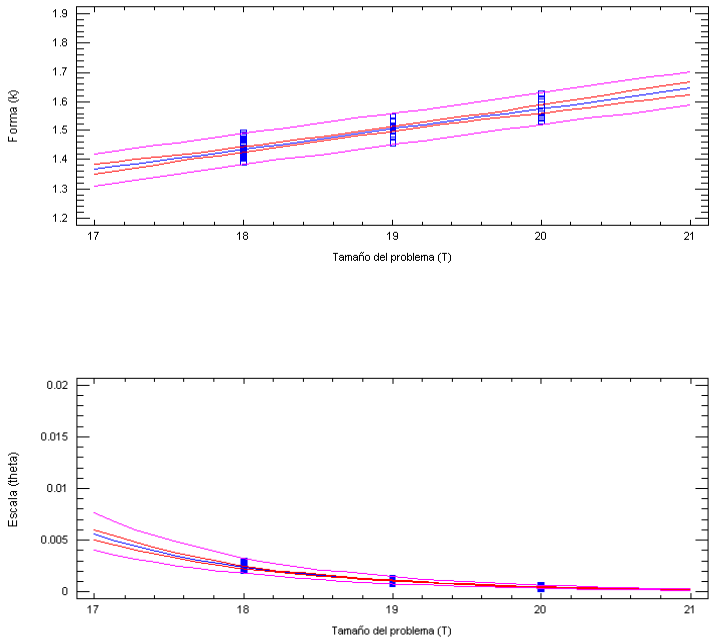

Figura 11: Modelo ajustado para los parámetros que constituyen el modelo de rendimiento Gamma de PBIL: forma (k) y escala (θ).

PBIL: *Modelo de rendimiento Gamma*	
Parámetro (P)	**Ecuación generalización**
Forma (k)	$\hat{k} = 0{,}190940 + 0{,}069186T;\ \forall T > 0;\ T \in N$
Escala (θ)	$\theta = e^{8{,}907130 - 0{,}829866T};\ \forall T > 0;\ T \in N$

Tabla 19: Generalización individual en función del tamaño del problema de los parámetros que definen el modelo de rendimiento Gamma del algoritmo PBIL.

mismos se realizará mediante la sustitución del tamaño correspondiente a la instancia que se desee resolver. Para conocer detalles acerca de la definición de intervalos de confianza de predicción véase la Sección 2.4.

6. Conclusiones

En el presente trabajo hemos obtenido una posible y simple generalización del rendimiento de los algoritmos CHC y PBIL en su aplicación al problema de la selección de la solución deseada en resolución de restricciones geométricas.

La predicción del rendimiento de los algoritmos a partir de la distribución Gamma lleva consigo una predicción para el conjunto de parámetros que influyen en la ejecución de CHC y PBIL. Para ello, hemos utilizado el modelo de regresión lineal simple estableciendo como variable regresora al tamaño del problema y como variables dependientes a todos y cada uno de los parámetros que definen el marco de ejecución y el modelo de rendimiento Gamma de los mencionados algoritmos.

Las hipótesis del modelo de regresión lineal simple y las restricciones para su aplicación se cumplen para la totalidad de los parámetros tanto a priori como a posteriori.

Las ecuaciones de generalización individual de los parámetros obtenidas respetan el dominio de valores de los mismos y siguen la evolución definida en la bibliografía en función de la complejidad del problema. Así mismo, se advierten variaciones coherentes de los valores de los parámetros ante el incremento del tamaño del problema.

7. Trabajos futuros

Como posibles trabajos futuros se proponen los siguientes:

- Verificación de la validez de las ecuaciones de generalización del rendimiento para la estimación del comportamiento de los algoritmos ante un conjunto de prueba constituido por instancias desconocidas del problema correspondientes a diferentes tamaños superiores a 2^{18}, 2^{19} y 2^{20}.
- Definición de un método alternativo de regresión conjunta para la predicción de valores de los parámetros en el que se considere la interrelación entre los parámetros que constituyen el marco de ejecución y la interrelación entre los parámetros que conforman el modelo de rendimiento Gamma.

8. Agradecimientos

Este trabajo ha sido realizado en el marco de los proyectos TIN2004-06326-C03-01 y TIN2004-06326-C03-02, financiados por FEDER y CICYT.

Referencias

[1] R. K. Ahuja and J. B. Orlin. Use of representative operation counts in computational testing of algorithms. *INFORMS Journal on Computing*, 8(3):318–330, 1996.

[2] J. Aldrich. R. A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, August 1997.

[3] M. M. Ali. Durbin-watson and generalized durbin-watson tests for autocorrelations and randomness. *Journal of Business and Economic Statistics*, 5(2):195–203, 1987.

[4] S. Baluja. Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical Report CMU-CS-94-163, Carnegie Mellon University, Pittsburgh, PA, 1994.

[5] W. Bouma, I. Fudos, C. Hoffman, J. Cai, and R. Paige. Geometric constraint solver. *Computer-Aided Design*, 27(6):487–501, June 1995.

[6] K. O. Bowman and L. R. Shenton. *Properties of Estimators for the Gamma Distribution.* CRC Press, 1988.

[7] G. C. Canavos. *Applied probability and statistical methods.* Addison-Wesley, 1984.

[8] G. M. Clarke and D. Cooke. *A Basic Course in Statistics.* Hodder Arnold, 1998.

[9] P. R. Cohen. *Empirical Methods for Artificial Intelligence.* MIT Press, 1995.

[10] J. L. Devore. *Probability and Statistics for Engineering and the Sciences.* Duxburg and Brooks Cole, Pacific Grove, CA, 6th edition, 2004.

[11] N. R. Draper and H. Smith. *Applied Regression Analysis.* John Wiley and Sons, Inc., New York, NY, 3rd edition, 1998.

[12] L. J. Eshelman. The CHC adaptative search algorithm: How to safe search when engaging in nontraditional genetic recombination. *Foundations of Genetic Algorithms I*, pages 265–283, 1991.

[13] J. Groβ. Linear regression. *Lecture Notes in Statistics*, 175(12), 2003.

[14] T. Hill and P. Lewicki. *Statistics: Methods and Applications.* Statsoft, Inc., Tulsa, OK, 2006.

[15] H. H. Hoos and T. Stützle. Evaluating las vegas algorithms – pitfalls and remedies. In *Proceedings of the 14th Conference on Uncertainly in Artificial Intelligence*, pages 238–245. Morgan Kaufmann, 1998.

[16] R. Joan-Arinyo, M. V. Luzón, and E. Yeguas. Ajuste, optimización y representación del rendimiento de PBIL y CHC aplicados al problema de la selección de la solución deseada. Technical Report LSI-07-41-R, Departament de Llenguatges i sistemes informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain, 2007.

[17] R. Joan-Arinyo, M. V. Luzón, and E. Yeguas. Comparación basada en distribuciones de longitud de tiempo de ejecución para PBIL y CHC aplicados al problema de la selección de la solución deseada. Technical Report LSI-07-37-R, Departament de Llenguatges i sistemes informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain, 2007.

[18] R. Joan-Arinyo, M. V. Luzón, and E. Yeguas. Parameter tuning for PBIL and CHC algorithms to solve the root identification problem in geometric constraint solving. Technical Report LSI-07-03-R, Departament de Llenguatges i sistemes informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain, 2007.

[19] K.-R. Koch. *Parameter Estimation and Hypothesis Testing in Linear Models.* Springer Verlag, Berlin; Heidelberg; New York, 1999.

[20] M. V. Luzón. *Resolución de Restricciones Geométricas. Selección de la Solución Deseada.* PhD thesis, Dpto. de Informática, Universidade de Vigo, Ourense, Spain, September 2001.

[21] E. A. Robinson. *Least Squares Regression Analysis in Terms of Linear Algebra.* Goose Pond Press, Houston, TX, 1981.

[22] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, MA, 1977.

[23] S. Weisberg. *Applied Linear Regression.* John Wiley and Sons, Inc., New York, NY, 3rd edition, March 2005.