Predicción de ingresos de causas penales mediante programación genética lineal.

Alberto David Garcete Rodríguez[1] y **Benjamín Barán**[2].
Facultad Politécnica, Universidad Nacional del Este.
Ciudad del Este, Paraguay.

[1]adgr_x@hotmail.com [2]bbaran@cba.com.py

Resumen

Este artículo propone una metodología de predicción de ingresos de causas penales utilizando una variación de la Programación Genética (GP): la Programación Genética Lineal (LGP). El estudio se realizó en base a datos mensuales recogidos durante siete años (2007 a 2013), provenientes de los siete Juzgados Penales de Garantías de Ciudad del Este. La verificación del método propuesto se llevó a cabo por comparación con modelos estadísticos, por lo que se ha estimado la misma serie de tiempo con estos modelos. La validación de los modelos fue realizada aplicando dos métricas; el error cuadrático medio y el error absoluto medio. El método LGP generó varios resultados debido a su capacidad de crear fórmulas matemáticas de manera aleatoria, de las mismas, tres resultados importantes fueron seleccionados para su comprobación. Dos resultados quedaron en primer y segundo lugar, mostrando mayor eficiencia respecto a los métodos estadísticos empleados. El tercer resultado seleccionado no obtuvo una buena predicción, pero sí un buen trazado en el gráfico, comparándolo con la serie de tiempo. En base a los resultados obtenidos, se concluye que los modelos generados por LGP son capaces de pronosticar con una buena precisión los ingresos penales, por encima de los métodos estadísticos empleados.
Descriptores: programación genética, programación genética lineal, ingresos penales.

Abstract

This paper proposes a methodology for predicting admissions of criminal cases using a variation of Genetic Programming (GP): Linear Genetic Programming (LGP). The study has been made based on monthly data collected during seven years (2007-2013) from the seven Criminal Courts of Ciudad del Este city. Verification of the proposed method was carried out by comparison with some statistical models. Time series has been estimated with these models. Validation of the models was made using two metrics; the mean square error and the mean absolute error. The LGP method generated various results from its capability to create random math formulas, from these, three important results were selected for testing. Two results ranked in first and second place, beating the statistical methods employed. The third selected result did not perform very well at predicting, but showed a good path in the graph compared to the time series. Conclusion based on the results obtained indicate that LGP generated models are able to predict criminal income with good accuracy achieving better results than the statistical methods employed.
Keywords: genetic programming, linear genetic programming, criminal income.

1. Introducción.

El aumento de delitos comunes en el entorno de la sociedad civil moviliza todo el aparato judicial en base a estos acontecimientos. Así, el inicio de juicios penales dentro de los estrados judiciales se ve afectado y existe un crecimiento progresivo año tras año [1, 2].

Investigar el aumento de delitos resulta interesante y beneficioso para mejorar la eficiencia del sistema judicial. Con este fin, se propone realizar predicciones utilizando un método denominado algoritmo evolutivo y métodos estadísticos, para de esta forma posibilitar la planificación de recursos necesarios para el Poder Judicial, sean estos humanos, espacio físico, insumos entre otros. El trabajo fue elaborado en base a datos que corresponden a un periodo de tiempo transcurrido entre enero/2007 y diciembre/2013. Se identificó la cantidad de delitos, luego se realizó una predicción de su crecimiento para intentar lograr una eficiente planificación de los recursos necesarios para el

cumplimiento de los fines específicos del Poder Judicial.

Se utilizó el método de desarrollo de programación genética lineal así como reconocidos métodos estadísticos, con el objetivo principal de hacer predicciones adecuadas a las necesidades de planificación.

1.1 Programación Genética.

La computación evolutiva es un conjunto de modelos computacionales, que desarrollan automáticamente programas de computadoras. Estos modelos computacionales operan por evolución de generaciones de las que son considerados como probables soluciones de un problema determinado [10].

La GP es una técnica de aprendizaje automático que se utiliza para optimizar una población de programas de acuerdo a la función de ajuste o aptitud (*fitness*), que sería empleada de evaluar la capacidad de cada programa.

La potencialidad de la programación genética reside en que posibilita desarrollar programas de forma automatizada. Su base biológica es la misma que los algoritmos genéticos, la diferencia radica en la forma en que se codifican los problemas.

1.2. Programación Genética Lineal.

Es una variación de la GP, que opera una secuencia de instrucciones imperativas de un lenguaje de programas de computadoras, incluyendo diversas formas de ejecución del programa [11]. En consecuencia la LGP es una evolución de la GP, donde el mayor cambio ocurre en la estructura de los programas. El tratamiento de estos programas en la GP es a través de estructura de árboles, sin embargo en la LGP los programas utilizan una estructura lineal.

La LGP se ha probado e interpretado generalmente en el lenguaje de programación de nivel medio C, sin embargo, está comprobado que los conceptos de la programación genética pueden ser traducidos en la mayoría de los lenguajes de programación imperativos modernos [10], e inclusive a nivel de lenguaje de máquina. En este trabajo, el algoritmo se ha desarrollado en el lenguaje de programación JAVA.

1.3. Serie temporal.

Una serie temporal es un conjunto de observaciones ordenadas en el tiempo, que pueden representar la evolución de una variable (económica, física, etc.) a lo largo de él [15]. El objetivo del análisis de una serie temporal es el conocimiento de su patrón de comportamiento, para así prever su evolución futura, suponiendo que las condiciones no variarán.

1.4. Estado del arte.

Como se dijera anteriormente, la Programación Genética Lineal es una extensión de la Programación Lineal. En la fuente de consulta [3], se presenta la LGP como una herramienta alternativa en la predicción de profundidad para tuberías sumergidas; los conjuntos de datos de las mediciones de laboratorio se obtuvieron de la literatura publicada y se utilizaron para desarrollar modelos LGP. El modelo LGP propuesto se comparó con sistemas adaptativos de inferencia *neuro-fuzzy* (ANFIS). Se observó que las predicciones de la LGP dieron buen resultado con los datos medidos; y bastante mejor que la ANFIS y la ecuación de regresión.

Los datos climáticos diarios, la temperatura del aire, la radiación solar, la velocidad del viento, la presión y la humedad de las tres estaciones meteorológicas automáticas de Fresno, Los Ángeles y San Diego, en California, fueron utilizados en [10] como datos para la LGP para estimar la evaporación. Las estimaciones LGP fueron comparadas con las de la programación de expresión génica (GEP), que es otra rama de la GP, perceptrones multicapa (MLP), redes neuronales de base radial (RBNN), regresión generalizada redes neuronales (GRNN) y *Stephens-Stewart*. Las actuaciones de los modelos fueron evaluados utilizando raíz media de errores cuadrados (RMSE), error absoluto medio (MAE) y el coeficiente estadístico de determinación (R2). Sobre la base de las comparaciones, se encontró que la técnica LGP se podría emplear con éxito en el modelado del proceso de evaporación a partir de los datos climáticos disponibles.

En [5] se observó un estudio del mercado de comercio de divisas (**Forex**), donde se ha utilizado un sistema de programación genética lineal para el mercado automatizado de *Forex* sobre cuatro pares de divisas principales. Se consideraron funciones de *fitness* con diferentes grados de conservadurismo a través de la incorporación de la máxima pérdida. Se ha examinado el uso de los tipos de acondicionamiento físico en el sistema de LGP de diferentes tendencias de valor de moneda en términos de rendimiento en el tiempo, que subyace a las estrategias comerciales y la rentabilidad global. Mediante un análisis de la rentabilidad del comercio se ha mostrado que el sistema de LGP es muy preciso, tanto en la compra para lograr ganancias como en la venta para evitar pérdidas, con niveles moderados de actividad comercial.

La estimación correcta de la concentración de sedimentos en suspensión transportada por un río es muy importante para muchos proyectos de recursos hídricos. En el trabajo [6], se propuso la aplicación de la programación genética lineal para la estimación de la concentración de sedimento en suspensión. La LGP se comparó con las redes neu-

rales adaptativas, *neuro-fuzzy* y modelos de curvas de calificación. El caudal diario y datos de concentración de sedimentos en suspensión a partir de dos estaciones, la estación de Río Valenciano y la estación de Quebrada Blanca, operado por el *US Geological Survey* (USGS); fueron utilizados como estudios de caso. La raíz media de los errores cuadrados (RMSE) y el coeficiente estadístico de determinación (R2), han sido utilizados para evaluar la exactitud de los modelos. La comparación de los resultados indicó que la LGP se comporta mejor que las redes neuronales, *neuro-fuzzy* y modelos de curvas de calificación.

Surcos laterales son estructuras de derivación utilizados ampliamente en el riego. En estas estructuras, la protección contra inundaciones y sistemas de alcantarillado combinado son necesarios. Este tema es abordado en el estudio [7]. La estimación precisa del coeficiente de descarga (Cd) de los vertederos laterales es esencial para calcular el perfil de la superficie del agua a través de estos y para determinar la tasa de flujo de salida lateral del sistema. En este trabajo, se utilizó una técnica de programación genética lineal para desarrollar nuevas fórmulas empíricas para la estimación de la Cd de los vertederos laterales rectangulares con aristas ubicadas en canales circulares. Se han empleado un total de 1.686 observaciones experimentales de laboratorio en ambos regímenes de sub y superflujo críticos con el fin de capacitar y validar los modelos propuestos. El rendimiento de los modelos basados en LGP se compararon con los de diferentes modelos de regresión lineal múltiple. Para determinar la eficiencia de los modelos se utilizaron la raíz media de los errores cuadrados, el error absoluto medio y el coeficiente de determinación. Los resultados indicaron en forma explícita, que el modelo basado en LGP utilizando funciones matemáticas se podría emplear con éxito en la estimación de Cd en ambas condiciones de flujo: sub y super críticas.

En el estudio [8], las técnicas de programación basada en árbol genético (TGP) y sus variantes recientes, programación genética lineal (LGP) y la programación genética de la expresión (GEP), han sido utilizadas para desarrollar nuevas ecuaciones de predicción para la capacidad de elevación de las cámaras de succión. La capacidad de elevación se formula en términos de variables flexibles. Se empleó una base de datos experimentales obtenidos a partir de la literatura para desarrollar los modelos. Además, se realizó un análisis estadístico convencional como referencia de los modelos propuestos. El análisis de sensibilidad y de parámetros se llevó a cabo para verificar los resultados. Fueron aplicados los modelos TGP, LGP y GEP por ser métodos eficaces para evaluar la capacidad de elevación horizontal, vertical e inclinada de cámaras a succión; estos alcanzaron un rendimiento de predicción mucho mejor comparados con los modelos que se encuentran en la literatura.

En el artículo [9], la programación genética lineal se utilizó para predecir la radiación solar global. La radiación solar se formula en términos de varios parámetros climatológicos y meteorológicos. Para desarrollar los modelos se han utilizado bases de datos completas que contienen datos mensuales recogidos durante 6 años (1995-2000) en dos ciudades nominales en Irán. Se establecieron modelos separados para cada ciudad. Para verificar el funcionamiento de los modelos propuestos, estos se han aplicado para estimar la radiación solar global de los datos de prueba de la base de datos. Se ha evaluado la contribución de los parámetros que afectan a la radiación solar a través de un análisis de sensibilidad. Los resultados indicaron que los modelos LGP dan estimaciones precisas de la radiación solar global y superan significativamente al modelo tradicional de Angstrom.

2. Metodología.

2.1. Algoritmo LGP Utilizado.

Algoritmo de Programación Genética Lineal.

01: Inicializar aleatoriamente una población de programas.

02: Aleatoriamente seleccionar los programas de la población y compararlos de acuerdo a su *fitness*. La medida del *fitness* define la calidad del programa en el problema que se espera resolver con el algoritmo.

03: Modificar los programas seleccionados utilizando alguno de los siguientes operadores de variación:

- Reproducción. Copia un programa sin cambiarlo,
- Cruzamiento. Intercambia subestructuras entre dos programas,
- Mutación. Cambia un registro, un operador o una instrucción en un programa en una posición aleatoria.

04: Se construye una nueva población con los programas variados y se calcula el valor de *fitness* de los nuevos programas.
05: Si el criterio de fin no se cumple, volver al paso 02.
06: Parar. El programa con mejor *fitness* representa la mejor solución encontrada.

El algoritmo de LGP fue diseñado para satisfacer las necesidades del problema, que consisten en lograr una predicción de ingresos de causas penales. Para ello se ha cumplido con los pasos básicos de la LGP, definiendo cada uno de ellos de acuerdo a lo referido más arriba. Para poder implementar este trabajo se ha realizado una adaptación de un algoritmo genético genérico.

2.2. Características de la LGP.

- Material genético lineal. Es una de las reglas en los algoritmos genéticos, y en este caso viene acompañado en una estructura lineal de programas de computadora.

- Material genético de longitud variable. La LGP trabaja sobre material genético que puede variar de tamaño. Pero por razones prácticas, generalmente se implementan limitaciones en el crecimiento, trabajando libremente podría permitirse crecimientos considerables a partir de una generación original que se produce aleatoriamente.

- Material genético ejecutable. La LGP trata con la evolución directa de programas de computadora. En la mayoría de los casos el material genético que está evolucionando es ejecutable, esto quiere decir que las estructuras son interpretadas por un lenguaje de computación existente. En todos los casos hay un proceso de ejecución del material genético, con el fin de ver directamente el comportamiento de la función deseada, a partir de la cual se obtiene el valor de aptitud.

- Cruce que preserva la sintaxis. En la mayor parte de los casos, se definen los operadores de cruces, de manera que estos preserven la corrección sintáctica del programa que es el material genético, todo esto definido por el lenguaje que se escoge para su representación.

2.3. Pasos aplicados para la elaboración del LGP.

2.3.1. Conjunto de registros.

Consiste en identificar el conjunto de registros que corresponde a variables o valores constantes, se puede ver como las entradas al programa. El conjunto de registros (junto con el conjunto de funciones) conforma los componentes a partir de los cuales se trata de construir un programa de computadora que implementa LGP para solucionar el problema.

El número de registros totales se debe definir al inicio del programa, y generalmente para todos los problemas no debe ser muy extenso. Como máximo se utiliza la cantidad de 256 registros [11]. Para el problema se ha optado por utilizar la cantidad de 22 registros (ver Tabla 1).

Tabla 1. Registros definidos para la LGP.

Registro	Descripción
r[0]	registro de salida
r[13]	valor anterior $(x_i - 1)$
r[12]	$x_i - 2$
r[11]	$x_i - 3$
r[10]	$x_i - 4$
r[15]	número de predicción -i
r[14]	pronóstico anterior
r[0]-r[9]	registros variables que se inicializan con valor 1
r[10]-r[21]	Registros constantes
r[16]	π
r[17]	e
r[18]	Valores Aleat. con rango [-1 a 1]
r[19]	Valores Aleat. con rango [-10 a 10]
r[20]	Valores Aleat. con rango [-100 a 100]
r[21]	Valores Aleat. con rango [0 a 1]

La entrada del programa es una sola y se almacena en el r[1] y la salida en el registro se almacena en el r[0], si las entradas son más de una, se utilizan los registros adyacentes.

Los registros quedan establecidos de la siguiente forma "r[0] al r[21]", de los cuales 12 registros son constantes y 10 registros variables. Los registros r[10] al r[15] son de entrada, conteniendo información introducida desde la fuente de datos conocida, que es la serie de tiempo estudiada.

2.3.2. Conjunto de funciones.

Consiste en identificar el conjunto de funciones que se van a usar, para generar la expresión matemática que trata de satisfacer la muestra dada finita de datos. A cada función corresponde un número de parámetros y un tipo de datos. Cada programa de computadora es una composición de funciones y registros, obtenidos de sus respectivos conjuntos.

Como en la programación genética, en la programación genética lineal se deben definir las terminales y las operaciones. Las terminales son los valores enteros, reales, las entradas y otros introducidos por el usuario. Las operaciones o funciones son aritméticas, exponenciales, condicionales, trigonométricas y boleanas (Ver Tabla 2).

Tabla 2. Operaciones para la LGP.

Tipo de Operación	Notación General	Rango Input
Aritméticas	$r_i = r_j + r_k r_i = r_j - r_k$, $r_i = r_j * r_k r_i = r_j / r_k$	$r_i, r_j, r_k \in$ R
Exponenciales	$r_i = r_j \hat{} r_k r_i = log(r_j)$, $r_i = \lvert r_j \rvert$	$r_i, r_j, r_k \in$ R
Trigonométricas	$r_i = seno(r_j)$, $r_i = cos(r_j)$	$r_i, r_j \in$ R
Booleanas	$r_i = r_j$ and $r_k r_i = r_j$ or r_k	$r_i, r_j, r_k \in \beta$
Condicionales	if $(r_j <= r_k)$ if $(r_j >= r_k)$	$r_i, r_j \in$ R

Las operaciones seleccionadas para utilizarlas en el trabajo fueron las aritméticas, las exponenciales y las trigonométricas, totalizando la cantidad de nueve operaciones para la selección aleatoria por parte de cada individuo a ser ejecutado en el sistema de operaciones.

2.3.3. Valor de aptitud o fitness.

Cada programa de computadora es un individuo en la población, y se lo debe medir en términos de qué tan bien se comporta en el ambiente del problema particular, a esta medida se le conoce como medida de aptitud y la naturaleza de la medida de aptitud varía con el problema. La aptitud de un programa se mide ponderando el error que se produce al ejecutar ese programa, entonces el mejor programa de computadora debe tener un error cercano a cero. Los individuos de la población inicial o generación cero, generalmente tienen un valor de aptitud muy baja. Sin embargo, algunos programas de la población son más aptos que otros, estas diferencias en el comportamiento son las que se deben explotar y explorar seguidamente. El *fitness* del individuo es calculado como inversamente proporcional a una función de error sobre un conjunto de valores de entrada/salida, que es conocido como secuencia de entrenamiento.

La función de error más utilizada para problemas de aproximación es la suma de los errores cuadráticos. Esta ecuación fue utilizada en el trabajo para calcular el *fitness* de cada individuo, y se la definió como la suma de los cuadrados de los errores para N valores de la muestra.

$$e = \frac{1}{N} \sum_{i=0}^{N} (x_i - p_i)^2 \qquad (1)$$

donde
x_i es el valor real de la posición i,
p_i es el pronóstico para el periodo i.

2.3.4. Parámetros para controlar la ejecución del algoritmo.

Al efectuar las operaciones genéticas sobre la población con la que se trabaja en el momento, la población de hijos reemplaza a la generación anterior, se mide la aptitud de cada individuo de la nueva población, y este proceso se debe repetir por muchas generaciones. En consecuencia, se deben definir los siguientes parámetros que controlan la ejecución del LGP: el tamaño de la población; el número de generaciones; las probabilidades de aplicación de los operadores genéticos de cruce y mutación; y la estrategia de selección de los individuos para la nueva generación. Ver Tabla 3.

Tabla 3. Parámetros utilizados en la LGP.

Parámetro	Valor
Longitud máxima de cromosomas para los individuos de la población inicial.	50
Longitud mínima de cromosomas para los individuos de la población inicial.	3
Longitud máxima de cromosomas de los individuos.	200
Tamaño de la población.	150
Tamaño de la población elite.	30
Cantidad de individuos por cada selección.	2
Probabilidad de cruzamiento por cada selección.	100 %
Probabilidad de mutación por cada selección.	30 %
Probabilidad de intensidad de mutación por cada selección.	20 %
Cantidad de operadores.	9
Cantidad total de registros.	22
Cantidad de registros constantes.	12
Cantidad de registros randómicos.	4
Cantidad de datos para entrenamiento.	72
Cantidad de datos para validación del modelo.	12

2.3.5. Método de selección, y criterio de parada.

El SelectorRuleta, es un método de selección de los individuos [12], que utiliza el mismo *fitness* asignado a cada uno de ellos como su valor de aptitud en base al peso que tienen sobre la solución del problema. Al tener el individuo un mayor valor de *fitness*, tiene mayor probabilidad de ser seleccionado al ejecutarse esta forma de selección.

El mecanismo trabaja de la siguiente forma: se selecciona a los padres utilizando el selector ruleta que posibilita tomar en forma aleatoria a los individuos de la población, a los cuales se aplica los operadores evolutivos de reproducción, mutación y cruzamiento si así correspondiere, y cuando se completa la cantidad establecida para la población, los padres se descartan y sigue la siguiente

generación con la nueva población formada por los hijos.

Como se denota, el algoritmo tiene sus límites y criterios de parada a fin de que pueda trabajar en forma. Se limita inicializando las variables al principio del algoritmo, estableciendo la cantidad de individuos "m" y la de generaciones "n", donde el criterio de parada del algoritmo es el número máximo de generaciones establecidas.

2.3.6. Puntos importantes tenidos en cuenta durante la creación del programa

Los individuos son una solución posible [13], y en el algoritmo son representados como un conjunto de ins-trucciones que se ejecutan en forma secuencial, a los efectos de proporcionar una posible solución al problema.

Ejemplo de individuo:

$$r[3] = r[2] + r[18]$$
$$r[2] = sen[1]$$
$$r[1] = r[2] + r[3]$$
$$r[0] = r[1]/r[2]$$

La población es el conjunto de estos individuos y en el trabajo se utilizó el sistema de generaciones, donde las nuevas poblaciones se construyen en base a la anterior, seleccionando a los padres para aplicarles los operadores evolutivos y generando con ellos los nuevos individuos.

La población inicial se crea al principio del algoritmo y se inicializa con individuos generados en forma aleatoria [14] utilizando como base los parámetros definidos en la Tabla 3. Esta población como se pudo observar, en cada ejecución inicial del programa genera individuos con fitness muy bajos para la solución del problema, y generación tras generación va mejorando por evolución.

Con el criterio de parada establecido y la población inicial generada, se asigna el fitness a cada individuo. Posteriormente se hace la selección a cuáles se aplicarán los operadores evolutivos de reproducción, cruce y mutación si así correspondiere. Generación tras generación a través del valor de aptitud se verifica cada individuo y se selecciona al mejor de cada generación para así obtener al mejor individuo como resultado del problema planteado.

En el operador evolutivo de reproducción, el individuo seleccionado pasa automáticamente a hacer parte de la siguiente población. En caso que el individuo posea mejor *fitness* que el individuo que actualmente se encuentre registrado como mejor individuo, se mantiene intacta su estructura.

Con el cruce, se seleccionan dos individuos de la población, se divide a cada individuo en sectores, y se realiza el cruce intercambiando los sectores seleccionados para intercambiar, generando así los nuevos individuos para la siguiente generación.

La mutación ocurre de forma diferente. En el algoritmo están establecidos ciertos métodos de mutación que se generan en forma aleatoria. Se establece un porcentaje de posibles mutaciones, aparte existe un criterio de cuánto por ciento de ese individuo puede ser mutado, estableciendo una cantidad fija. Asimismo para cada mutación que se realiza al individuo, el mismo puede cambiar dependiendo de cuál sea el caso, una operación, un registro o todo un cromosoma.

2.3.7. Serie de tiempo estudiada en esta investigación.

Dado que no se trata de fenómenos deterministas, sino sujetos a una aleatoriedad, el estudio del comportamiento pasado ayuda a inferir la estructura que posibilite predecir su comportamiento futuro, pero es necesaria una gran cautela en la previsión debido a la inestabilidad que generalmente tienen estos modelos. La particular forma de la información disponible de una serie cronológica (se dispone de datos en periodos regulares de tiempo) hace que las técnicas habituales de inferencia estadística no sean válidas para estos casos, ya que se encuentran "N" muestras de tamaño "i" procedentes de otras tantas poblaciones de características y distribución desconocidas. Ver Fig. 1.

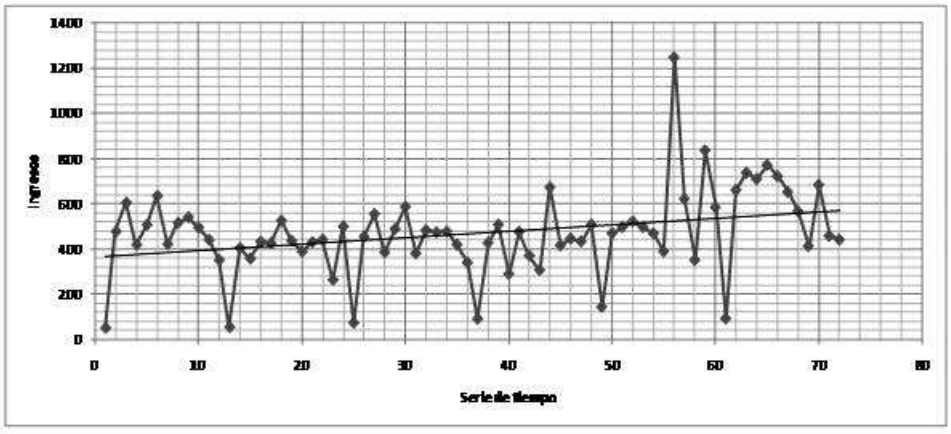

Figura 1. Serie de tiempo de ingresos de causas penales.

2.3.8. Métodos estadísticos.

En la planificación del trabajo se ha establecido que los datos de estudio también deberían ser aplicados a metodologías estadísticas; en este caso a métodos de serie de tiempo y suavizado, contra los que competiría el LGP. Fueron aplicadas las fórmulas de series de tiempo que se detallan a continuación.

Regresión lineal. En la regresión lineal simple, existe una relación entre la variable X (independiente) y la variable Y (dependiente), que puede representarse mediante una línea recta (Ec. 2).

$$y = mx + b \tag{2}$$

Los valores de "b" y "m" se eligen de manera que se minimice la suma de las distancias cuadráticas entre la línea de regresión y los puntos de datos.

Promedio Móvil. Es el movimiento medio de orden "N" de una serie de valores o es el promedio aritmético de las "N" observaciones más recientes tal como se observa en su fórmula (Ec. 3)..

$$S_t = \frac{1}{N} \sum_{i=t}^{t-N+1} D_i \tag{3}$$

Utilizando adecuadamente los movimientos medios se eliminan las variaciones estacionales, cíclicas e irregulares, quedando sólo el movimiento de tendencia. El inconveniente de este modelo es la pérdida de datos iniciales de la serie original. Otra situación que también se observa es que a medida que crece "N", la cantidad de datos nuevos se reduce.

Suavizado Exponencial. Es uno de los métodos más utilizados entre las series de tiempo estacionarias. Contiene un mecanismo de corrección que ajusta los pronósticos en dirección opuesta a los errores pasados, donde las ponderaciones disminuyen exponencialmente. Se emplea con la fórmula que sigue (4).

$$F_t = \alpha D_{t-1} + (1 - \alpha) F_{t-1} \tag{4}$$

donde:
α es la constante de suavización que pondera relativamente la observación de demanda actual,
$1 - \alpha$ es el peso asignado a la observación de demanda pasada.

Suavizado Exponencial con tendencia. Este método requiere de la especificación de dos constantes de suavización:
α: que suaviza el valor de la serie (promedio - estacionario),
β: que suaviza la tendencia (pendiente de los datos).

Las constantes de suavización pueden ser las mismas; sin embargo en la mayoría de las aplicaciones se da mayor estabilidad a la estimación de la pendiente que a la de la constante, es decir:

$$\beta \leq \alpha \tag{5}$$

Por otra parte, deben utilizarse dos parámetros para la estimación del pronóstico, estos son:
S_o: corte de la recta de regresión b
G_o: pendiente de la recta de regresión m
Las ecuaciones aplicadas a este método son:

$$S_t = \alpha D_{t-1} + (1 - \alpha)(S_{t-1} + T_{t-1}) \tag{6}$$

$$G_t = \beta(S_t - S_{t-1}) + (1 - \beta) T_{t-1} \tag{7}$$

$$F_t = S_t - T_t \tag{8}$$

2.4. Datos y métodos aplicados.

Las pruebas experimentales fueron realizadas con los modelos expuestos anteriormente. Al principio se contaba con 72 meses de datos históricos de los ingresos de causas penales de los siete juzgados penales de garantías de Ciudad del Este, del periodo de tiempo de enero del 2007 a diciembre del 2012 extraídos del sitio web de la Excma. Corte Suprema de Justicia [1, 2]. Sin embargo, con el correr del tiempo durante el desarrollo de esta investigación, se ha llegado a obtener de la misma fuente de información, datos del año 2013, los cuales fueron agregados a la presente investigación a los efectos de expandir la cantidad de datos para mejorar el despliegue de los métodos aplicados.

2.4.1. Duistribución de datos.

Las pruebas se realizaron con la distribución de los datos establecidos de la siguiente manera:
Entrenamiento o prueba: 72 meses.
Validación o Pronóstico: 12 meses.
Total de datos disponibles: 84 meses (enero/2007 a diciembre/2013)

2.4.2. Métricas de desempeño.

Las evaluaciones de las metodologías implementadas se realizaron por métricas de desempeño. Se evaluaron en base a los errores generados por cada una de ellas, con los resultados obtenidos durante el periodo de validación en el caso del LGP y de pronóstico en el caso de los métodos estadísticos.

Primera métrica La fórmula del error cuadrático medio aplicado a los planteamientos es la siguiente: el error es inversamente proporcional a la suma del cuadrado de la diferencia entre el valor

real y de pronóstico de cada periodo, considerando las "N" muestras del periodo de estudio.

$$e = \frac{1}{N}\sum_{i=1}^{N}(x_i - p_i)^2 \tag{9}$$

Segunda métrica. La fórmula del error absoluto medio es: el error es inversamente proporcional a la suma de la diferencia absoluta entre el valor real y el pronóstico de cada periodo, considerando las "N" muestras del periodo en estudio.

$$e = \frac{1}{N}\sum_{i=1}^{N}|x_i - p_i| \tag{10}$$

2.4.3. Procedimiento de ejecución de cada método.

Para todos los métodos LGP y Series de Tiempo:

A: Se utilizaron los 84 meses históricos de la serie de tiempo en estudio, divididos en dos partes, una de 72 meses para el entrenamiento o prueba y otra de 12 meses para la validación o pronóstico.

B: Los registros mencionados, se aplicaron a cada uno de los métodos para obtener los resultados. En el caso del LGP, la aplicación se realizó con el sistema desarrollado y en el caso de los métodos de series de tiempo, la aplicación se realizó con cada una de sus fórmulas correspondientes.

C: A los resultados obtenidos en los 12 pronósticos realizados por cada método, se le aplicaron los cálculos de las métricas de los errores seleccionados, a los efectos de obtener la información necesaria para la medición y evaluación que se realizó en las comparaciones pertinentes.

D: Por último, con los valores obtenidos de las métricas de cada método ejecutado, se procedió a la evaluación de cada uno de ellos, para obtener las conclusiones del trabajo.

Procedimiento exclusivo para el método LGP:

A: La población es de estado generacional, donde los padres son reemplazados por los hijos en la siguiente generación.

B: En cada generación el sistema realiza las selecciones de los programas a través de la ruleta que tiene como método de selección y medida el *fitness* de cada uno de ellos, cuanto mayor *fitness* tienen mayor es la probabilidad de ser seleccionados. Los seleccionados son indicados para que se apliquen a ellos los operadores evolutivos y así generar los nuevos individuos de la siguiente población.

C: En cada generación se evalúa el mejor individuo entre los generados (Ver Ec. 1), para que al finalizar la ejecución del programa, éste pueda tener como uno de los resultados al mejor individuo de la población considerando sólo los datos de entrenamiento, que será la solución al problema planteado.

D: El pronóstico de los 12 meses correspondientes al año 2013 fue generado mediante la ejecución de la fórmula matemática obtenida con el mejor individuo seleccionado por el LGP.

3. Resultados.

El análisis de los resultados fue realizado por medio de tablas que dan un mejor panorama para observar las soluciones. Se procedió en base a las métricas establecidas donde cada tipo de error genera un resultado diferente que se puede observar en cada solución tomada como fuente de información final.

3.1. Reglas obtenidas con el LGP.

La aleatoriedad del LGP para obtener soluciones distintas en cada ejecución de programa, ha generado buenos resultados en la mayoría de las ejecuciones que fueron realizadas.

Las ejecuciones de programa contenían variaciones de datos de entrada, registros, variables, tamaño de programa; a los efectos de encontrar la solución óptima para resolver el problema, véase Tabla 1. En cada prueba además de encontrarse buenos resultados se observó que estas soluciones simulan y logran representar las funciones de los métodos estadísticos estudiados en el trabajo. Para este estudio se apartaron estas tres reglas obtenidas por el LGP para su demostración, véase la tabla 4.

Tabla 4. Mejores individuos de las ejecuciones del LGP.

Individuo A	Individuo B	Individuo C
$r[9] = r[15] - r[21]$	$r[6] = \|r[17]\|$	$r[4] = r[14] + r[6]$
$r[0] = r[7] + r[0]$	$r[6] = r[18]^{r[3]}$	$r[9] = r[6] + r[4]$
$r[1] = r[15]/r[19]$	$r[3] = sen[12]$	$r[5] = r[7] + r[17]$
$r[7] = r[1] + r[1]$	$r[6] = r[16] * r[11]$	$r[3] = r[13] + r[10]$
$r[0] = \|r[9]\|$	$r[6] = r[17] - r[21]$	$r[3] = sen[15]$
$r[8] = r[21] + r[0]$	$r[2] = r[5] - r[7]$	$r[5] = r[7] + r[17]$
$r[8] = r[20] + r[0]$	$r[6] = cosen[9]$	$r[4] = \|r[10]\|$
$r[0] = \|r[9]\|$	$r[0] = \|r[9]\|$	$r[0] = r[13]^{r[5]}$
$r[0] = r[8] + r[1]$	$r[3] = sen[18]$	$r[8] = r[8]/r[7]$
$r[0] = r[7] + r[0]$	$r[3] = ln[14]$	$r[0] = r[7] * r[2]$
$r[8] = r[21] + r[0]$	$r[4] = ln[6]$	$r[0] = r[1] - r[0]$
$r[0] = r[7] + r[0]$	$r[9] = sen[13]$	$r[4] = cosen[5]$
$r[8] = r[21] + r[0]$	$r[9] = ln[16]$	$r[6] = r[13] - r[1]$
$r[0] = r[7] + r[0]$	$r[5] = r[10] * r[7]$	$r[7] = r[16] + r[16]$
$r[8] = r[21] + r[0]$	$r[5] = \|r[16]\|$	$r[3] = cosen[16]$
$r[0] = r[8] + r[1]$	$r[0] = cosen[11]$	$r[4] = r[4] * r[6]$
$r[7] = r[1] + r[1]$	$r[5] = cosen[19]$	$r[0] = r[7] * r[2]$
$r[0] = r[7] + r[0]$	$r[4] = \|r[15]\|$	$r[4] = r[14] + r[6]$
$r[8] = r[21] + r[0]$	$r[4] = \|r[19]\|$	$r[5] = r[18] + r[16]$
$r[0] = r[19] * r[8]$	$r[8] = r[14] * r[20]$	$r[7] = r[16] + r[16]$
	$r[2] = sen[21]$	$r[0] = r[19] + r[12]$
	$r[7] = r[3]/r[5]$	
	$r[4] = sen[14]$	
	$r[4] = r[21] - r[15]$	
	$r[2] = r[7]/r[4]$	
	$r[8] = r[18] - r[5]$	
	$r[7] = ln[3]$	
	$r[3] = r[9]/r[16]$	
	$r[0] = r[2] - r[5]$	
	$r[3] = sen[0]$	
	$r[8] = r[16] + r[3]$	
	$r[0] = ln[7]$	
	$r[7] = \|r[12]\|$	
	$r[7] = r[8] + r[9]$	
	$r[0] = r[7] * r[20]$	

3.2. Comparación y evaluación de resultados.

Los métodos de LGP y los métodos estadísticos aplicados en este trabajo han arrojado resultados muy interesantes. El objetivo principal del trabajo fue ejecutar cada una de estas metodologías y comparar las soluciones que se obtuvieron en base a la serie de tiempo de los ingresos de causas penales de los Juzgados Penales de Garantías de Ciudad del Este. La comparación se realiza a continuación.

3.3. Primera métrica del Error Cuadrático Medio.

Para observar mejor los resultados obtenidos con la métrica del error cuadrático medio, se ha montado la tabla 5 donde las soluciones de cada una de las técnicas ejecutadas se han asentado.

Tabla 5. Comparación de resultados con error cuadrático medio.

Mes	e^2PGL(A)	e^2PGL(B)	e^2(S.Exp.)	e^2(S.Exp.Tend.)	e^2(Regresión)	e^2(P. móvil.)	e^2PGL(C)
Enero	14361,47	35628,71	67667,37	52617,80	59413,70	81035,11	47552,66
Febrero	3764,75	6809,11	6573,60	316,19	28828,59	3927,11	15892,57
Marzo	19565,15	9093,21	24134,67	47484,53	1100,22	35344,00	76139,73
Abril	10631,37	408,67	3506,06	7544,58	3,50	16129,00	27867,05
Mayo	1002,25	7740,42	9307,00	6562,36	18201,12	7511,11	28583,22
Junio	4022,77	15009,00	7548,77	3351,48	27209,20	16469,44	26917,56
Julio	33926,80	65216,27	28944,29	17540,12	80651,84	32881,78	22519,72
Agosto	17413,24	34560,86	2216,35	98,19	52915,29	1111,11	4364,68
Septiembre	32860,20	15342,73	82801,86	116682,34	7212,46	98177,78	135375,64
Octubre	461,95	6413,29	747,92	658,03	13482,55	841,00	12754,15
Noviembre	4321,74	71,97	5415,65	6725,72	737,39	3844,00	12783,86
Diciembre	2118,48	8396,44	4204,92	4523,74	18822,61	13378,78	486,90
Total error:	144450,19	204690,68	243068,46	264105,08	308578,47	310650,22	411237,72
Error medio:	12037,52	17057,56	20255,70	22008,76	25714,87	25887,52	34269,81

Analizando la tabla 5, el resultado final del error sitúa en primer y segundo lugar con los mejores resultados al método LGP con los Individuos A y B, con un total de 12.037,52 y 17.057,56 puntos respectivamente, quedando en tercer lugar el método de suavizado exponencial con un total de 20.255,70 puntos y en cuarto lugar el método suavizado exponencial con tendencia con un total de 22.008,76 puntos.

Los errores obtenidos con la métrica del error cuadrático medio, muestran que los Individuos A y B del método LGP tuvieron un comportamiento relativamente bueno y constante, siempre quedando cada resultado entre los mejores lugares, o sea con errores bajos en comparación con los demás métodos.

3.3. Segunda métrica del error absoluto medio.

La métrica del error medio absoluto da la opción de verificar si la técnica que resulta con mejores resultados en la primera métrica, es la misma o varía con este otro modo de observar el error que tiene cada solución.

Observando la tabla 6, se nota que el método que genera los mejores resultados quedando en primer lugar es el LGP con sus Individuos A y B, con un error total de 95,83 y 111,89 puntos respectivamente. Asimismo en los siguientes lugares sí hubo variación comparando con la anterior

métrica; en este caso el tercer lugar queda a favor del método de suavizado exponencial con tendencia con un error total de 112,47 puntos y el cuarto lugar en base a los mejores resultados queda se sitúa el método de suavizado exponencial con error total de 117,49 puntos.

Tabla 6. Comparación de resultados con error absoluto medio.

Mes	$\|e\|$**PGL(A)**	$\|e\|$**PGL(B)**	$\|e\|$**(S.Exp.Tend.)**	$\|e\|$**(S.Exp.)**	$\|e\|$**(P. móvil.)**	$\|e\|$**(Regresión)**	$\|e\|$**PGL(C)**
Enero	119,84	188,76	229,39	260,13	284,67	243,75	218,07
Febrero	61,36	82,52	17,78	81,08	62,67	169,79	126,07
Marzo	139,88	95,36	217,91	155,35	188,00	33,17	275,93
Abril	103,11	20,22	86,86	59,21	127,00	1,87	166,93
Mayo	31,66	87,98	81,01	96,47	86,67	134,91	169,07
Junio	63,43	122,51	57,89	86,88	128,33	164,95	164,07
Julio	184,19	255,37	132,44	170,13	181,33	283,99	150,07
Agosto	131,96	185,91	9,91	47,08	33,33	230,03	66,07
Septiembre	181,27	123,87	341,59	287,75	313,33	84,93	367,93
Octubre	21,49	80,08	25,65	27,35	29,00	116,11	112,93
Noviembre	65,74	8,48	82,01	73,59	62,00	27,15	113,07
Diciembre	46,03	91,63	67,26	64,85	115,67	137,20	22,07
Total error:	1149,95	1342,68	1349,69	1409,88	1612,00	1627,86	1952,26
Error medio:	95,83	111,89	112,47	117,49	134,33	135,63	162,69

Con el error absoluto medio se nota que los errores generados por el programa LGP con sus Individuos A y B, mantienen un promedio bajo en todos los puntos validados, lo cual es importante destacar, ya que a consecuencia de ser constante en sus errores, sus resultados son siempre cercanos a las demandas estudiadas, y esto hace que el método tenga buena probabilidad de éxito en sus predicciones.

El programa LGP presentó en dos casos los mejores resultados para la serie de tiempo de ingresos de causas penales de Juzgados de Garantías de Ciudad del Este, esto comprueba la eficiencia de la metodología con estas pruebas realizadas.

El método LGP tiene la capacidad de generar funciones matemáticas aleatorias. En este caso se ha comprobado que debido a su aleatoriedad, los resultados no siempre serán buenos, como se pudo observar en el Individuo C, a través de las métricas establecidas. Sin embargo, no todo se echa a perder con este individuo, ya que fue capaz de generar un buen recorrido gráfico en comparación con la serie de tiempo estudiada. Entonces se puede deducir que el LGP es capaz de obtener resultados muy buenos para series de tiempo no lineales o irregulares, véase la figura 2.

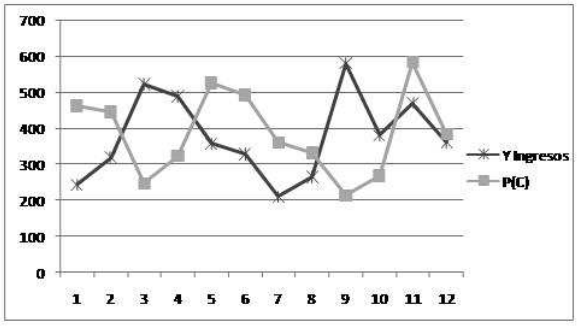

Figura 2. Predicción ejecución 24 del programa LGP.

4. Conclusión.

La LGP fue ejecutada varias veces y en la mayoría de las ejecuciones generaba resultados muy buenos. Observando estos resultados y comparándolos con cada modelo matemático, se notó que las soluciones simulan y logran representar a las soluciones obtenidas por estos métodos matemáticos, con la gran diferencia que el LGP logra generalmente mejores resultados.

La LGP es apta para ser aplicada como metodología de predicción de los ingresos de causas penales de los Juzgados de Garantías de Ciudad del Este, por la eficiencia obtenida en sus resultados en base a esta serie de tiempo estudiada en este trabajo, inclusive pudiendo extenderse su aplicación a Juzgados de otras regiones del país.

El éxito del LGP radica en que los resultados que genera se mantienen con un promedio de error bajo, quedando siempre entre los primeros lugares en cada predicción realizada, y al efectuar las sumatorias de los errores y aplicar las métricas asignadas, no varía sus resultados para ninguna de ellas. Sin embargo como se vio entre los resultados matemáticos, hubo cambio de posición entre

los dos mejores resultados entre cada métrica, esto obedece a que sus resultados generan en ciertos puntos de predicción errores muy altos que afectan su desempeño general.

Las variantes que se pueden aplicar a la LGP para su ejecución posibilitan obtener resultados para modelos de predicciones lineales y no lineales, otro de los motivos por lo que genera soluciones eficientes en comparación a los métodos estadísticos clásicos que demuestran limitaciones para predicciones de series de tiempos no lineales [16].

Por último, observando el crecimiento de los delitos y así como se puede ver el incremento en la estructura y de los RR.HH. de la Excma. Corte Suprema de Justicia [1] y [2], entonces se puede concluir que con este tipo de investigaciones, se puede realizar una reingeniería de recursos, llevando infraestructura y recurso humano donde más se necesite. En consecuencia aplicando este tipo de técnicas y tecnologías se obtendría mejores resultados y ajustando los procesos se podrá ofrecer un mejor servicio en el área afectada: la justicia.

Trabajos futuros.

Los trabajos futuros que se podrían emprender en la línea de investigación del presente trabajo incluyen los siguientes:

1. Análisis de parámetros de inicialización y los valores de entrada.
2. Comparación de la LGP con otros modelos no lineales, como las redes neuronales.
3. Estudio de problemas con múltiples salidas.
4. Implementación de la LGP con enfoques multi-objetivo.

Referencias bibliográficas

[1] Excma. Corte Suprema de Justicia. Disponible: `http://www.Csj.Gov.Py/` Acceso: 2014.

[2] Poder Judicial. Disponible en: `http://www.Pj.Gov.Py/` Acceso: circa 2014.

[3] Azamathulla, H. M., Guven, A. & Demir, Y. K. Linear genetic programming to scour below submerged pipeline. OceanEngineering, 38(8), pp. 995-1000, 2011.

[4] Koza, J. R. Introduction to genetic programming tutorial: from the basics to human-competitive results. In Proceedings of the 12th annual conference companion on Genetic and evolutionary computation. ACM., pp. 2137-2262, 2010.

[5] Wilson, G. & Banzhaf, W. Inter-day foreign exchange trading using linear genetic programming. In Proceedings of the 12th annual conference on Genetic and evolutionary computation. ACM, pp. 1139-1146, 2010.

[6] Kisi, O. & Guven, A. A machine code-based genetic programming for suspended sediment concentration estimation. Advances in Engineering Software, 41(7), 939-945, 2010.

[7] Uyumaz, A., Mehr, A. D., Kahya, E. & Erdem, H. Rectangular side weirs discharge coefficient estimation in circular channels using linear genetic programming approach. Journal of Hydroinformatics Vol 16 No 6 pp 1318?1330 © IWA DOI de la publicación: 10.2166/hydro.2014.112, 2014.

[8] Alavi, A. H., Aminian, P., Gandomi, A. H. & Esmaeili, M. A. Genetic-based modeling of uplift capacity of suction caissons. Expert Systems with Applications, 38(10), 12608-12618. 2011.

[9] Shavandi, H. & Ramyani, S. S. A linear genetic programming approach for the prediction of solar global radiation. Neural Computing and Applications, 23(3-4), 1197-1204, 2013.

[10] Koza, J. R. Human-competitive results produced by genetic programming. Genetic Programming and Evolvable Machines, 11(3-4), 251-284, 2010.

[11] Sanchez, R. "Aplicación de la programación genética lineal para la predicción de indicadores macro económicos", Proyecto final de tesis, Facultad de Ciencias y Tecnología Universidad Católica Nuestra Señora De La Asunción, 2008.

[12] Koza, J. R. "Herarchical Genetic Algorithms Operating On Populations Of Computer Programs". En N. S. Sridharan, Editor, Proceedings Of 11th International Joint Conference On Artificial Intelligence, San Mateo, Morgan Kaufmann, California, 1989.

[13] Koza, J. R. Introduction to genetic programming tutorial: from the basics to human-competitive results. En Proceedings of the 12th annual conference companion on Genetic and evolutionary computation (pp. 2137-2262). ACM, (julio de 2010).

[14] Poli, R. y Koza, J. Genetic Programming (pp. 143-185). Springer US, 2014.

[15] Jenkins, M. "Time Series Analysis Forecasting And Control", 2da Edition, Holden-Day, San Francisco, 1976.

[16] Dabhi, V. y Chaudhary, S. Time Series Modeling and Prediction Using Postfix Genetic Programming. Advanced Computing & Communication Technologies (ACCT), 2014 IEEE Fourth International Conference(pp. 307-314).