UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE PREDICCIÓN DE DEMANDA DE LA POBLACIÓN PENAL A TRAVÉS DE MINERÍA DE DATOS Y DINÁMICA DE SISTEMAS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PABLO ANDRÉS LEMUS HENRÍQUEZ

PROFESOR GUÍA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
CARLOS REVECO DÍAZ
VLADIMIR MARIANOV KLUGE
DANIEL PAVEZ TRALMA

SANTIAGO, CHILE
2014

A mis profesores por el apoyo tanto académico como anímico entregado.

A mi familia por acompañarme en todo este proceso convertirme en un profesional.

Y a mi mujer, por convertirse en mi más grande soporte y hacerme sentir que nada es imposible.

Resumen Ejecutivo

El problema de Sobrepoblación penal en las cárceles de Chile es una realidad que Gendarmería de Chile ha tenido que enfrentar en los últimos años. El tener una buena predicción de la población penal permite tener información con la cual tomar mejores decisiones (planificación de cárceles considerando la demanda futura por ejemplo) para enfrentar esta situación.

Con el fin de modelar este problema se utilizó la metodología KDD para la construcción de un modelo de flujos de la población penal con alcance regional (usando como caso de estudios la región metropolitana). Dada la existencia de variables estacionarias y no estacionarias, la naturaleza cíclica del problema y también el desarrollar un nuevo enfoque científico para la predicción es que se propone crear una herramienta híbrida de minería de datos (para predecir la variable no estacionaria) y simulación de dinámica de sistemas.

La variable no estacionaria (Aprehendidos Mayores de Edad) se predijo evaluando diversas técnicas de minería de datos, obteniéndose la mejor predicción con la técnica Support Vector Machine con Kernel Radial, la cual tuvo un error medio porcentual igual a 4,53%.

Luego, se simuló el modelo mediante Dinámica de Sistemas y Eventos Discretos, comparando los escenarios con y sin reincidencia. Se obtuvo mejores resultados con la simulación de dinámica de sistema considerando la reincidencia, con un error medio porcentual en la predicción con horizonte de 1 año menor a un 4% para la población de Condenados, y menor a un 2% para la población de Imputados.

Se evaluaron las siguientes estrategias para disminuir el tamaño de la población penal: Reducir en un 5% la reincidencia y deportar 500 Condenados de nacionalidad extranjera a sus países de origen. Donde la primera estrategia muestra un resultado con resultados entre corto y mediano plazo pero que se estabiliza para el mediano plazo; y la segunda estrategia muestra una solución instantánea pero que se estabiliza en el mediano plazo, y que entrega resultados menos óptimos que la primera estrategia para el mediano plazo.

Los resultados muestran el considerar la reincidencia al predecir la población penal se obtienen mejores resultados, Siendo posible generar una herramienta flexible, capaz de ser remodelada y servir de utilidad para diversas instituciones, en función de generar estrategias óptimas y efectivas en torno a la disminución de la sobrepoblación penal.

Tabla De Contenido

Índice de ilustraciones

Índice de tablas

Capítulo 1

1 Introducción

El año 2005 se implementó la reforma procesal penal en la región metropolitana de Chile, la cual cambió la constitución y el rigor de las leyes. Esto produjo un aumento en la demanda de la población penal a través del tiempo, aumentando el hacinamiento en las cárceles.

La sobrepoblación penal ha sido un problema para Gendarmería de Chile por los siguientes motivos:

- Dificulta la tarea de velar por el cumplimiento de las detenciones preventivas y condenas que dicten en los tribunales [17] ya que mientras más presos hayan más difícil es prevenir conductas de riesgo que pongan en riesgo el cumplimiento de esta tarea.
- La falta de infraestructura para satisfacer el tamaño de la población penal en Chile impide ofrecer un trato digno y respetar los derechos inalienables de las personas [17].
- Para satisfacer esta falta de infraestructura se recurre al arriendo de espacio en cárceles concesionadas [17], lo cual implica un gasto económico extra para el fisco.

El anticiparse a esta población en el futuro permitiría tomar decisiones como asignación de recursos, planificación de futuras cárceles, etc.

El análisis para predicción y pronóstico de series de tiempo se ha utilizado en diversos problemas, como predecir cambios climáticos, precios de acciones en el mercado financiero, el producto interno bruto, etc., utilizando información histórica de sí misma (valores de la variable a predecir en el pasado) como variables explicativas para predecir sus valores futuros.

Por otro lado está la simulación de Dinámica de Sistemas. Una técnica para analizar y modelar el comportamiento temporal de un sistema en entornos complejos, basada en modelamientos cíclicos entre elementos, y las demoras en la información y los materiales dentro del sistema.

Dada la existencia de factores no estacionarios (como el ingreso a las cárceles de los criminales), a que el proceso penal corresponde a un flujo cíclico de personas y, sumado al problema de hacinamiento, se propone la idea de proponer una herramienta híbrida que permita predecir la población penal y medir el impacto de posibles estrategias para reducir la sobrepoblación penal.

1.1. Alcance del trabajo

En esta tesis se desarrollará la unión de los dos enfoques mencionados anteriormente para concretar una herramienta, con la cual se va a predecir la población penal de la Región Metropolitana de Chile (específicamente la de Imputados y Condenados) a nivel mensual con un horizonte de 12 y 36 meses. Además se evaluará el impacto en la demanda de la población penal que causarían las siguientes estrategias: Reducir la reincidencia en un 5% y la deportación de condenados extranjeros.

1.2. Contexto

Gendarmería de Chile es un servicio público dependiente del Ministerio de Justicia, que tiene por finalidad atender, vigilar y contribuir a la reinserción social de las personas que por resolución de autoridades competentes, fueren detenidas o privadas de libertad y cumplir las demás funciones que le señala la ley [10].

Esta institución busca cumplir 3 objetivos estratégicos [17]:

- Garantizar el cumplimiento eficaz de la detención preventiva y de las condenas que los Tribunales determinen, previniendo conductas y situaciones que pongan en riesgo el cumplimiento de este objetivo.
- Proporcionar una atención y un trato digno a la población puesta bajo la custodia del Servicio, reconociendo y respetando los derechos inalienables de las personas, entregando un conjunto de condiciones básicas de vida que faciliten el ejercicio de los derechos no restringidos por la reclusión.
- Fomentar conductas, habilidades, destrezas y capacidades que incrementen las probabilidades de reinserción social de la población penal, involucrando en este proceso a sus familias, instituciones, empresas y comunidad en general.

Para cumplir estos objetivos, Gendarmería de Chile debe asignar bien sus recursos (Gendarmes, diseño y construcción de cárceles, etc…), porque de lo contrario podrían ocurrir consecuencias graves en cada recinto penitenciario (motines, peleas, etc.).

Definición 1.1 (Sobrepoblación penal): Se define sobrepoblación penal como la cantidad de presos en que se supera la capacidad de una Unidad Penal [11]. Por ejemplo, sea X_a la capacidad de una Unidad Penal “a”, e Y el tamaño de la población penal de la misma Unidad Penal, el índice de sobrepoblación de la Unidad Penal “a” se define de la siguiente forma:

$$\acute{I}ndiceSobrepoblación_a = \begin{cases} 0 & si \quad X_a > Y_a \\ \frac{Y_a}{X_a} - 1 & si \quad \sim \end{cases} \tag{1.1}$$

Definición 1.2 (Hacinamiento): Se define hacinamiento como la cantidad de presos en que se supera el doble de la capacidad de una Unidad Penal [11]. Por ejemplo, sea X_a la capacidad de una Unidad Penal "a", e Y el tamaño de la población penal de la misma Unidad Penal, el índice de sobrepoblación de la Unidad Penal "a" se define de la siguiente forma:

$$Hacinamiento_a = \begin{cases} 0 & si \quad X_a > Y_a \\ \frac{Y_a}{X_a} - 1{,}2 & si \quad \sim \end{cases} \tag{1.1}$$

Las celdas contienen a más internos de la capacidad para la cual fueron diseñadas (llegando incluso a convivir 16 internos en celdas de 3 x 2 metros en el año 2011 cuando llegó a existir un sobre poblamiento de un 60% en las cárceles de Chile [41] cuando debería haber un preso cada 4 metros cuadrados según el Comité Europeo para la Prevención de las Torturas y de las Penas o Tratos Inhumanos y Degradantes [33]*),* por lo que no se cumpliría con el segundo objetivo estratégico porque los presos han vivido en condiciones indignas para cualquier persona. Y como el personal capacitado (gendarmes) no es suficiente para hacerse cargo de los internos, tanto sus vidas como la de los presos corren peligro si llega a ocurrir un evento (pelea, motín, etc.) porque al no poder contener la situación podría ocurrir una tragedia. Un claro ejemplo de esto fue la tragedia del incendio de la cárcel de San Miguel [12] ocurrida el 8 de Diciembre del año 2010, donde murieron 81 reos dado que solo estaban resguardados por 5 gendarmes que no pudieron evacuar a todos los internos de sus celdas.

Región/Establecimientos	SUBSISTEMA CERRADO			CAPACIDAD SEGÚN DISEÑO			INDICE DE USO DE CAPACIDAD		
	Hombres	Mujeres	Total	Hombre	Mujer	Total	Hombre	Mujer	Total
DE ARICA Y PARINACOTA	1.617	244	1.861	1.448	504	1.952	113,1%	49,6%	96,7%
DE TARAPACA	2.085	279	2.364	2.225	464	2.689	94,0%	60,1%	88,1%
DE ANTOFAGASTA	2.019	223	2.242	1.072	174	1.246	189,4%	133,3%	181,5%
DE ATACAMA	869	97	966	446	68	514	202,5%	147,1%	195,1%
DE COQUIMBO	2.088	152	2.240	2.140	133	2.273	99,9%	115,8%	100,8%
DE VALPARAISO	4.368	387	4.755	2.890	288	3.178	156,7%	139,9%	155,2%
DE O'HIGGINS	2.291	149	2.440	2.232	157	2.389	104,5%	96,8%	104,0%
DEL MAULE	1.861	90	1.951	1.270	58	1.328	153,0%	167,2%	153,6%
DEL BIO BIO	3.692	209	3.901	3.783	194	3.977	98,1%	107,7%	98,6%
DE LA ARAUCANIA	1.618	74	1.692	1.518	98	1.616	107,0%	75,5%	105,1%
DE LOS RIOS	1.035	58	1.093	1.498	61	1.559	70,8%	95,1%	71,8%
DE LOS LAGOS	1.615	62	1.677	1.932	60	1.992	86,9%	106,7%	87,4%
DE AYSEN	207	4	211	242	14	256	89,3%	28,6%	85,9%
DE MAGALLANES Y ANTARTICA	249	7	256	443	42	485	58,7%	16,7%	55,1%
METROPOLITANA	14.358	1.257	15.615	10.786	1.848	12.634	136,7%	68,0%	126,7%
TOTAL NACIONAL	39.972	3.292	43.264	33.925	4.163	38.088	120,6%	80,2%	116,1%

Tabla 1. Población penal, capacidad e índice de ocupación de Unidades Penales de cada Región para el día 31 de Marzo de 2014 [17]

De la tabla anterior se observa que hay varias regiones que están en un estado de Sobrepoblación Penal, y en caso particular la región de Atacama se encuentra en un nivel de Hacinamiento de un 82,5% para la población penal masculina.

Si bien se ha logrado disminuir el índice de ocupación de las Unidades Penales entre el año 2010 al año 2013 a un 22% [2], y finalmente el año 2014 a un 20%, esto sigue siendo una cifra relevante si a esto le sumamos que el hacinamiento significa una pérdida económica para el fisco porque para cada empresa que sea dueña de una cárcel concesionada que tenga una población mayor a un 120% de su capacidad el fisco debe compensar con 100UTM ($4.180.000) por cada día que se supere ese porcentaje [23]. De hecho, desde el año 2008 hasta Mayo del año 2012 el fisco ha pagado más de $5.159 millones de peso [23].

1.3. Series de Tiempo versus Dinámica de Sistemas

Las series de tiempo permiten predecir el valor que una variable toma en un momento determinado del tiempo al recoger información sobre su evolución a lo largo del tiempo y explotar el patrón de regularidad que muestran los datos.

Para construir un modelo de serie de tiempo lo único que se necesita es la información muestral de la variable a analizar. Esto hace que sea más fácil de estimar debido a que no siempre se dispone de los datos de las variables exógenas (variables explicativas) y no se necesita predecir los valores de las variables exógenas para poder predecir la variable que nos interesa.

Sin embargo su gran falencia es que no permiten realizar un análisis con respecto a las causas que explican el valor de esa variable.

La Dinámica de Sistemas es una técnica para analizar y modelar el comportamiento temporal de un sistema en entornos complejos.

A diferencia de las series de tiempo, su objetivo básico es llegar a comprender las causas estructurales que provocan el comportamiento del sistema. Esto implica aumentar el conocimiento de cada elemento y ver como diferentes acciones involucradas acentúan o atenúan las tendencias de comportamiento implícitas dentro de su sistema.

El problema de esta técnica es que se deben conocer a priori el comportamiento de las variables participantes y contar con información real de cada una de ellas, lo cual aumenta la complejidad de llevarla a cabo.

Al mezclar las series de tiempo con la dinámica de sistemas se obtiene una herramienta que disminuye el costo de predecir las variables explicativas (ya que para aquellas variables estables se podría utilizar el promedio, mientras que para las que tengan un comportamiento inestable en el tiempo) y a la vez permite analizar cómo estas afectan a la variable que queremos predecir.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar una herramienta híbrida de minería de datos y dinámica de sistemas que permita predecir la demanda de la población penal de la región metropolitana de Chile y obtener información relevante con respecto a los efectos de estrategias de confrontación al problema del hacinamiento en las cárceles.

1.4.2. Objetivos específicos

a) Entender y estructurar el proceso penal.
b) Comprender la participación de Gendarmería de Chile dentro del proceso penal y sus posibles acciones a realizar.
c) Modelar el problema en función de los enfoques propuestos anteriormente y de la información recopilada.
d) Analizar y predecir la serie de tiempo de la(s) variable(s) no estacionaria(s) del modelo.
e) Determinar los valores de las variables y parámetros restantes del modelo.
f) Simular el modelo y realizar una validación de los resultados.
g) Obtener una predicción en un horizonte de 12 con un error medio porcentual menor o igual a un 4%
h) Obtener una predicción en un horizonte de 36 meses de plazo.
i) Analizar el impacto de posibles estrategias en la demanda de la población penal.

1.5. Metodología

- Analizar datos disponibles de las siguientes instituciones: Carabineros de Chile y Gendarmería de Chile.
- Estudiar literatura relacionada con el problema y modelar el problema según enfoques propuestos.
- Evaluar distintos métodos de predicción de series de tiempo para pronosticar los valores de la(s) variable(s) no estacionarias y determinar el mejor método a utilizar.
- Determinar los valores de las variables y parámetros restantes del modelo en función de sus valores históricos.
- Simular el problema con técnicas de dinámicas de sistemas y de eventos discretos, validar los resultados y determinar la mejor técnica a utilizar.
- Predecir la demanda de población de Condenados e Imputados en un horizonte de 12 y 36 meses de plazo.
- Evaluar el impacto en la demanda de la población penal que producirían las dos estrategias mencionadas anteriormente.
- Analizar resultados y concluir.

1.6. Resultados esperados

Esta tesis busca comprobar las siguientes hipótesis:

- Considerar la reincidencia dentro de modelos de predicción entrega mejores resultados que modelar sin tomarla en cuenta.
- La simulación de dinámica de sistemas para este problema en particular funciona mejor que la simulación de eventos discretos.
- Un mix de minería de datos para predicción de series de tiempo en conjunto con la simulación es una herramienta que genera resultados de confianza para predecir la demanda de la población penal de la región metropolitana.
- Reducir la reincidencia y deportar condenados extranjeros son estrategias de impacto sostenido a través del tiempo para combatir el hacinamiento en las cárceles.

1.7. Estructura del trabajo

Este trabajo está dividido en 2 partes. Los fundamentos teóricos (capítulos 2, 3 y 4) y el desarrollo de la metodología propuesta.

El capítulo 2 presenta los conceptos generales de la metodología de descubrimiento de conocimiento en bases de datos (KDD, Knowledge Discovery in Databases), fundamento base para el desarrollo de trabajos en minería de datos.

El capítulo 3 presenta los métodos para predecir la serie de tiempo de la variable Aprehendidos Mayores de Edad de la Región Metropolitana (variable no estacionaria con tendencia al alza dentro del problema modelado), centrándose en los conceptos y los parámetros de cada método que fueron modificados en el desarrollo de la tesis.

El capítulo 4 presenta los tipos de simulación existentes, los paradigmas de simulación más usados y la metodología para desarrollar un proyecto de simulación.

El capítulo 5 presenta el modelamiento del problema, desde el levantamiento de la información, pasando por el estudio de la literatura existente y los factores considerados para, finalmente, modelar el problema.

El capítulo 6 presenta el cálculo de los valores de cada componente del modelo, a excepción de la variable Aprehendidos Mayores de Edad.

El capítulo 7 presenta la predicción de la variable Aprehendidos Mayores de Edad mediante técnicas de minería de datos, desde el cálculo de los parámetros óptimos para cada técnica evaluada hasta la elección de la mejor técnica a utilizar.

El capítulo 8 presenta los resultados de la predicción de la población penal, evaluando los dos escenarios mencionados y las distintas técnicas de simulación mencionadas. Además se predice la población penal en un horizonte de 36 meses.

El capítulo 9 presenta el análisis de los efectos que tendrían en la demanda de la población penal las estrategias evaluadas para combatir el hacinamiento en las cárceles.

El capítulo 10 presenta las conclusiones de este estudio y las futuras líneas de investigación para continuar mejorando la herramienta desarrollada.

El anexo A presenta la información obtenida de cada fuente detallada en el capítulo 5.

Finalmente, el anexo B presenta el detalle del cálculo de cada variable del modelo según la metodología determinada en el capítulo 6.

PARTE I: TEORÍA Y MÉTODOS

Capítulo 2

2 Conceptos de minería de datos

La minería de datos ha se ha vuelto una de las áreas de estudio de mayor desarrollo e impacto en las distintas ciencias, principalmente porque presenta una oportunidad muy importante tanto para una empresa como para una institución: ¿Qué hacer con los datos existentes? Acorde a Myatt [31]: "el volumen de datos generado ha llevado a una sobrecarga de información y la habilidad de obtener algún sentido de ella se ha vuelto cada vez más importante". En este contexto se define la minería de datos como "la extracción de información previamente desconocida de grandes bases de datos que pueden ruidosas y tener datos perdidos" [9].

Este campo ha tenido un auge rápido y masivo, por lo que se ha desarrollado un marco teórico amplio y detallado donde se cuentan con todos los pasos a seguir para que al desarrollar un nuevo proyecto se cuente con una mayor probabilidad de éxito.

Este capítulo describirá un resumen de los distintos puntos que han expresado los autores en la literatura y que son necesarios para extraer de forma rigurosa información de una base de datos.

En particular, se estudia el camino a seguir para definir el problema, luego cada uno de los puntos de la metodología KDD para crear el modelo y finalmente los requerimientos mínimos que deben tener las conclusiones de un estudio de minería de datos.

2.1. Pasos a definir para modelar el problema a resolver

Todo estudio de minería de datos aborda un problema particular para una entidad, y en este caso es la sobrepoblación y el hacinamiento en las cárceles de Chile,

Sin embargo, en la literatura existen pocas publicaciones que noten los pasos que se deben seguir para asegurar que el problema esté bien definido. Mackinson y Glick [25] definen los siguientes puntos que se deben considerar antes de construir un modelo para resolver el problema:

- **Objetivos**

  Después de haber identificado el problema a resolver hay que definir el objetivo general que se desea perseguir, como ejemplifica el planteado en este trabajo: “se desea construir una herramienta híbrida mezclando la minería de datos con simulación de dinámica de sistemas que prediga la demanda de la población penal de la región metropolitana de Chile”. Este objetivo debe ser dividido en objetivos específicos, identificando los distintos puntos donde se desea tener un resultado. En este paso es recomendable definir un criterio de éxito cuantificable, como por ejemplo, uno de los objetivos específicos de esta tesis es “que el error medio porcentual en un horizonte de un año sea menor a un 4%”.

- **Entregables**

  Corresponde al acuerdo entre las partes interesadas sobre el proyecto a realizar donde se definen las expectativas a cumplir, el alcance de los resultados a obtener y la validación de los resultados (pues no sirve perder tres meses en desarrollar un modelo con un 95% de eficacia si se puede desarrollar uno con un 85% en días y esto suple los problemas de la institución). También se debe tener en cuenta el tiempo del proyecto, el costo de las equivocaciones y el equipo de trabajo necesario para cumplir con los objetivos.

  También es importante definir el formato de entrega de los resultados, como por ejemplo un software, un modelo, etc.

  En este caso el entregable corresponde a esta tesis (donde se buscan cumplir los objetivos determinados en el capítulo 1) y la herramienta computacional de predicción de la población penal.

- **Roles y responsabilidades**

  La gran mayoría de los proyectos corporativos se desarrollan en equipos de trabajos (más de un integrante) donde se definen los roles y responsabilidades de cada uno de sus integrantes dependiendo de su experiencia y conocimientos para obtener un desarrollo más técnico y específico, como también una colaboración para realizar mejor las tareas designadas y en menos tiempo.
  Cabe señalar que quien desarrolla el proyecto y quien se beneficia de él no siempre son la misma persona porque el proyecto puede ser externalizado.

En este caso el beneficiario de este proyecto es Gendarmería de Chile y el proyecto fue externalizado a un alumno tesista de la Universidad de Chile, quien a su vez fue supervisado por sus profesores.

- **Costos y Carta Gantt**

  Los plazos y costos deben ser definidos por ambas partes (ejecutor y cliente) para asegurar que los intereses de cada una sean resueltos (el equipo ejecutor desearía tener mucho tiempo, pero el consumidor desea tenerlo lo más rápido posible).

  Es recomendable reajustar la carta Gantt al completar cada hito del proyecto para reflejar los retrasos o adelantos y las nuevas metas de tiempo. Con todo lo anterior se puede desarrollar una presupuesto general del proyecto y evaluar finalmente si el proyecto es rentable.

  Una vez concretado todo esto se puede proceder con la creación del modelo.

  En esta ocasión el plazo de realización del proyecto estuvo dado por los plazos de realización de tesis estipulados para los alumnos del Magíster en Gestión de Operaciones de la Universidad de Chile.

2.2. Desarrollo del modelo: Metodología KDD

La metodología KDD [13,25] es el resultado de la estandarización de los procesos relacionados con la transformación de grandes volúmenes de datos en conocimiento útil en cualquier área del conocimiento que necesite de modelos matemáticos (generalmente estadísticos) para interpretar relaciones no triviales en bases de datos. Esquemáticamente el proceso se puede observar en la ilustración 1.

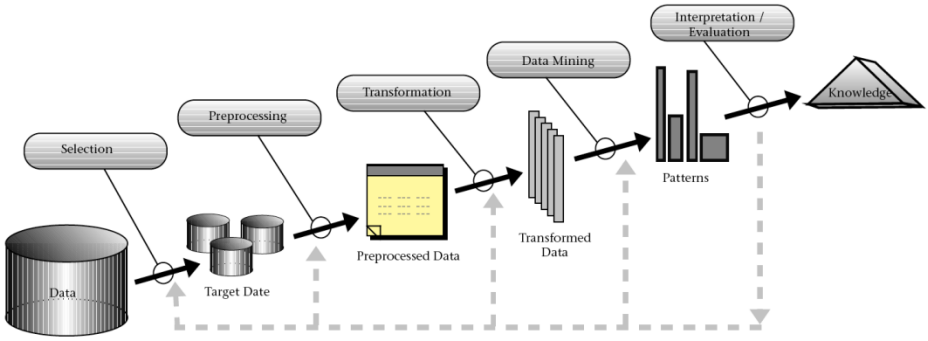

Ilustración 1. Metodología KDD [6]

2.2.1. Selección de Datos

El primer paso de la metodología KDD es la selección de las distintas fuentes de información, sean estas externas (como la estadística de la cantidad de aprehendidos mayores de edad de la base de datos de Carabineros de Chile), bases de datos corporativas o data warehouses, o creada específicamente para este problema (a través de encuestas o indicadores calculados a partir de otras variables) para poder obtener los datos necesarios para desarrollar el modelo de forma correcta. Teniendo en cuenta que la calidad de los datos es por mucho el factor más importante para la creación de un modelo, ya que deben representar patrones en los datos fidedignamente para entregar una base de datos para resolver el problema. Es por eso que existe un alto riesgo de introducir errores en esta etapa, porque afecta la validez del modelo desarrollado.

Este paso se desarrolla orientado a la determinación de la variable a predecir y a las variables independientes.

2.2.2. Preprocesamiento de Datos

Esta etapa es la que suele tomar la mayor parte del tiempo del proyecto ya que al recopilar información de distintas fuentes, estas pueden estar duplicadas, incompletas y/o con inconsistencias.

Tareas comunes que deben realizarse son:

- **Descripción previa de los datos**

  Este paso consiste en comprender la calidad de los datos antes de trabajar con ellos. Estos deben ser examinados mediante herramientas estadísticas clásicas, como frecuencias y medias.

- **Datos Nulos**

  Los registros con datos nulos son un problema común y grave a la hora de realizar un modelo de minería de datos ya que el único modelo que puede utilizar los datos nulos son los árboles de decisión.

  Para esto se han desarrollado diversas técnicas para el tratamiento de datos nulos:
  - Eliminar registros con datos nulos.
  - Completar datos de manera aleatoria (según distribución de la variable).
  - Completar datos con el valor de la media.
  - Métodos de expectación – maximización.
  - Regresiones.
  - Etc.

- **Limpieza de datos**

  Se deben explorar las bases para estandarizar algunas entradas y eliminar los datos que no corresponden al atributo. Estos pueden incluir diversas formas de registrar algún valor pues por ejemplo un atributo “región” puede incorporar “región metropolitana”, “RM”, “Reg. Metrop.”, etc. O, pueden figurar valores inconsistentes, como edades negativas y otros errores.

- **Incorporación de información conocida adicional**

  Este paso busca complementar los datos para poder consolidar o eliminar algunos atributos, sin embargo ay que tener cuidado debido a que se pueden alterar sin querer el valor de algunos campos como también duplicar datos o generar datos inconsistentes.

La salida de este paso es una base de datos limpia, completa y preparada para ser adaptada para su uso en un modelo estadístico.

En el caso de esta tesis se realizaron estos pasos con las bases de datos detalladas en el capítulo 5.

2.2.3. Transformación de Datos

Los modelos de minería de datos suelen tener limitaciones en cuanto a los tipos de datos que aceptan, y es en esta fase donde se debe decidir cuál es la transformación de cada atributo que satisface de mejor manera las necesidades del modelador.

Algunas transformaciones potenciales corresponden a:

- **Normalización**

  Los atributos continuos (como la edad o el ingreso por ejemplo) pueden ser escalados a un intervalo debido a que la magnitud de las diferencias de escala entre las variables independientes puede afectarla precisión del modelo predictivo a elaborar.

  Algunos métodos corresponden a la normalización en el intervalo [0,1], el escalamiento z (restar la media de la variable y dividir por la desviación estándar observada) y el escalamiento decimal, que corresponde a dividir por 10^n donde n es la cantidad de dígitos que posee el valor más alto de la variable.

  En el capítulo 5 se mostrará la normalización de algunas variables.

- **Mapeo de variables ordinales**

  Las variables que representen categorías con un ordenamiento definido pueden ser transformadas en variables numéricas definiendo alguna magnitud explícita para cada categoría, representando así una distancia conceptual en términos de una distancia numérica. El riesgo que se corre es que la definición y la interpretación de esta distancia numérica es subjetiva y por ende puede llevar a introducir patrones donde realmente no los hay.

- **Transformación de variables categóricas**

  Las variables categóricas se pueden representar matemáticamente en base a transformar un atributo con n categorías en n-1 variables tales que si un registro particular pertenece a la categoría i entonces la variable i toma valor uno y el resto

toma valor 0. Una de estas categorías se fija como categoría de referencia, pues en caso contrario existirá un problema de multicolinealidad en las variables[1] y no se le debe crear variable asociada, al ser determinable a partir de las demás.

En el capítulo 7 se mostrará un ejemplo de esto.

- **Discretización de variables continuas**

  En el caso de que los valores de una variable se encuentren concentrados en intervalos disjuntos de valores se puede construir una nueva variable que categorice estos rangos. De esta forma se puede capturar de manera más "limpia" las relaciones que un modelo puede interpretar de forma errónea o simplemente pasar por alto.

  Otro uso es cuando se está en presencia de una variable con comportamiento no lineal y se está utilizando un modelo lineal. Una opción es eliminar la variable con comportamiento no lineal y generar nuevas variables categóricas, una para cada tramo en el cual se observe comportamiento lineal.

- **Agregación y cambio de rango**

  Las variables pueden ser transformadas para aumentar su capacidad discriminante cuando están altamente agregadas a través de cambiar el rango. Esto puede ser realizado por ejemplo calculando el logaritmo natural de alguna de ellas, ralentizando el crecimiento de la variable.

  Otra opción es construir una nueva variable calculando el total de un conjunto de variables afines, como por ejemplo calcular el ingreso total por familia para un conjunto de habitantes sumando sobre los ingresos de cada integrante.

Una vez transformado los datos se dispone de una base de datos en condiciones óptimas para aplicar un método de minería de datos. Esta base de datos contiene en el óptimo todos los registros con valores confiables y dentro de rango, escalados o transformados para asegurar que sean discriminantes y representen el problema.

---

[1] *Dos o más variables son colineales si una variable se puede expresar como una función lineal de las demás. En este caso, si se crearan n variables binarias entonces una de ellas (arbitraria) se podría representar como* $x_i = 1 - \sum_{j \neq i} x_j$.

2.2.4. Minería de Datos

El siguiente corresponde a la iteración del modelo definido con diferentes métodos de minería de datos para encontrar el de mejor funcionamiento para el problema particular. Donde previamente se debe dividir la base de datos para validar los resultados de forma óptima, por lo que se debe definir una parte como base de entrenamiento de los modelos y otra parte como base de validación o de test, siendo esta última utilizada para la comparación de resultados obtenidos con la base de entrenamiento. Se recomienda que este conjunto de validación sea entre el 20% y el 30% del total de registros, utilizándose en promedio un 25% [19].
El conjunto de datos restantes debe ser nuevamente dividido para realizar análisis de validación cruzada del modelo, analizando así la estabilidad de los resultados; por lo general se busca dividir el conjunto en particiones iguales, volviendo a dejar afuera el 20% o 25% de la muestra total y entrenando con el 60% o 50% restante y luego rotando el conjunto que queda fuera, sin nunca incorporar el conjunto de test separado en el paso previo.

Luego de realizar esta división se debe aplicar el modelo de minería de datos, definiendo claramente una metodología de prueba para los distintos factores que influyen en un modelo y que por lo menos incluya los siguientes pasos:

- **Re-selección de atributos**

  Si bien anteriormente se seleccionaron los datos pertinentes para definir los atributos del modelo, no todos estos son discriminantes, e incluso si lo fuesen no lo serán en la misma medida. Este paso comúnmente es ignorado en los modelos debido a que al aplicar las técnicas de modelación se suele creer que estos se seleccionaran solos al otorgar pesos bajos a los valores de estos atributos [45].

  Para esto existen varios métodos desarrollados para seleccionar atributos, y el encontrar uno que se adapte mejor a la técnica a utilizar queda a criterio del desarrollador.

- **Selección del modelo a utilizar o a medir eficacia**

  El siguiente factor a considerar corresponde a la técnica a utilizar. Acorde al problema particular existen una variedad de modelos que sirven para abordarlo, todos con distintos niveles de complejidad, tiempo de computación, conocimiento necesario por parte del experto, etcétera. La selección de los modelos debe pasar por considerar el nivel de conocimiento que el experto tenga de los mismos, la complejidad necesaria

para resolver el problema (si es necesario un complejo modelo no lineal o basta con otro más sencillo), las herramientas computacionales disponibles y por último la complejidad en el tiempo de desarrollo.

- **Selección de parámetros del modelo**

  Cada modelo posee una variedad de parámetros que deben ser ajustados y cuyos valores depende fuertemente del problema particular a ser abordado.
  Se debe definir una metodología de prueba para determinar el mejor modelo con el fin de minimizar el riesgo de fallar en encontrar la mejor combinación de parámetros, la cual puede diferir de la metodología aplicada para determinar el valor de los parámetros de otro modelo.

Una vez diseñados todos estos pasos se deben probar los modelos y validar los resultados en base a la capacidad discriminante que posea (medida en el conjunto de testeo) y probada su estabilidad considerando las distintas combinaciones de las particiones del conjunto de entrenamiento.

La salida de este paso es un modelo completamente validado que debe ser evaluado en el paso siguiente para su implementación.

El desarrollo de esta unidad se puede apreciar en el capítulo 7 de esta tesis.

2.2.5. Interpretación y Evaluación

Una vez obtenidos los resultados del modelo, el siguiente paso es crear nuevo conocimiento a partir de ellos, analizando el resumen de la información, y diseñando los entregables y/o implementación.

- **Entregables**

  En esta fase se deben diseñar los entregables del proyecto, siendo los más populares los reportes, los módulos para incrustarse a otro software y los softwares autónomos. Los reportes sirven para describir los descubrimientos realizados y las acciones a seguir para la implementación, por lo que siempre debiesen estar presentes en un proyecto de minería de datos ya que permiten tener la información teórica y práctica resultante disponible para consultas posteriores. Los módulos de software para integración en sistemas existentes tienen la ventaja de ser efectivos en costos, requieren bajo nivel de

entrenamiento en el personal y se puede acceder de forma rápida a la información existente. El software autónomo crea una solución que puede ser de despliegue rápido al venir “listo para ser usado”, pero su integración en los sistemas de una empresa puede ser mucho más difícil.
En esta ocasión los entregables corresponden a la herramienta autónoma propuesta como objetivo general de este estudio y los conocimientos obtenidos descritos en esta tesis.

- **Implementación**

  En esta fase se debe planificar y ejecutar la implementación, en base a definir participaciones dentro de esta, definir responsabilidades (si es el caso), identificar capacitaciones necesarias y discutir la metodología para mantener actualizados los modelos.

  Otro paso corresponde a describir como se medirá la efectividad de los modelos descritos y también su nivel de ajuste a los datos a lo largo del tiempo.

  En esta ocasión corresponde a la capacitación del personal de Gendarmería de Chile a cargo de utilizar esta herramienta, desde la captura y procesamiento de datos hasta la obtención del pronóstico de la población penal.

El último paso de la metodología KDD es la comprensión de los resultados del proyecto. El esfuerzo realizado y el tiempo invertido para obtener un modelo estadístico que extraiga conocimiento de bases de datos no es despreciable, por lo que se debe aprovechar al máximo todas sus potencialidades.

Por lo general en el desarrollo de un modelo de minería de datos se obtiene mucho conocimiento externo al que aporta el modelo en sí mismo, pues si se sigue la metodología KDD entonces en cada paso se va alcanzando un nivel nuevo de conocimiento de las bases de datos y de las características de las mismas (ya que las bases de dato son el fiel reflejo del significado de los procesos de la entidad que los desarrolla). Es por eso que nunca se debe perder de vista los objetivos a cumplir para poder llegar a una correcta y efectiva comprensión de los resultados.

Capítulo 3

3 Técnicas de minería de datos aplicadas a predecir series de tiempo

Una serie de tiempo es una secuencia de datos medidos en determinados momentos del tiempo, ordenados cronológicamente, y normalmente espaciados entre si uniformemente.

El análisis de estas series tiene diversos usos, tanto para extraer información representativa como también la posibilidad de extrapolar y predecir su comportamiento futuro, siendo este último uno de los más habituales. Algunos ejemplos de predicción de series de tiempo (con las técnicas a describir en este capítulo) son: predecir el precio del cobre [14], demanda de telefonía móvil [21], índices del retail u otros índices financieros [18], variaciones de costos de insumos para proyectos de construcción [22], etc…

En términos más específicos, las series temporales se usan para estudiar la relación causal entre diversas variables que cambian con el tiempo y se influyen entre sí. Desde el punto de vista probabilístico una serie temporal es una sucesión de variables aleatorias indexadas según parámetro creciente (o decreciente) en el tiempo. Cuando la esperanza matemática de dichas variables aleatorias no es constante, ni varía de manera cíclica, se dice que la serie no es estacionaria y presenta una tendencia secular.

A continuación se detallan las técnicas de minería de datos utilizadas para predecir los valores futuros de una serie de tiempo desarrollada en esta tesis.

3.1. W-Linear Regression

Esta técnica consiste en una regresión lineal en la cual el modelo (selección de variables) se determina mediante el criterio Akaike [5].

3.1.1. Regresión Lineal

La regresión lineal [38] es un método matemático que modela la relación entre una variable dependiente Y, las variables independientes X_i, parámetros β_i y un término aleatorio ε.

Este modelo puede ser expresado como: $Y = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot X_i + \varepsilon$

Cabe destacar que el término β_0 es conocido como la intersección o término constante.

La primera forma de regresiones lineales documentada fue el método de los mínimos cuadrados, el cual fue publicado por Legendre en 1805, y en dónde se concluía una versión del teorema de Gauss-Márkov.

3.1.2. Criterio Akaike

El Criterio de Información Akaike [5] o Akaike Informarion criterion (AIC) en inglés, es una medida de la calidad relativa del modelo estadístico para un conjunto dado de datos. Este criterio muestra un equilibrio entre la complejidad del modelo y la bondad de ajuste de este.

AIC se basa en la entropía de la información: se ofrece una estimación relativa de la información perdida cuando se utiliza un modelo determinado para representar el proceso que realmente genera los datos.

El índice AIC se define de la siguiente forma: $AIC = 2 \cdot k - 2 \cdot \ln(L)$

Donde k es el número de parámetros en el modelo y L es la función de máxima verosimilitud del modelo estimado. Y el modelo elegido es aquél que posea el menor valor de AIC

De esto observamos que no solo recompensa el ajuste de bondad obtenido por la función de máxima verosimilitud, sino que también penaliza la complejidad del modelo, es decir, mientras más variables se consideren mayor es la penalización, por lo que se disminuye el número de parámetros libres que aumentarían el ajuste de bondad en la función de máxima verosimilitud.

La gracia de este criterio es que si bien no se puede elegir con certeza un modelo estadístico de predicción, se puede estimar, a través de AIC, el mejor modelo predictivo entre los modelos candidatos, en el sentido que proporciona la aproximación más cercana a la realidad o al verdadero modelo. Además, su simplicidad y facilidad para ser implementado, y el hecho de

que no existe el problema de especificar subjetivamente un nivel de significancia arbitrario para contrastar dos modelos lo hacen un criterio muy utilizado en la práctica.

3.2. Redes Neuronales

Esta técnica tiene su inspiración en las neuronas humanas, las cuales consisten de un soma, o cuerpo celular, que posee todos los elementos (núcleo, ribosomas, etc.) que la capacitan para sintetizar proteínas y neurotransmisores; el axón, en cuyo terminal se fijan las sinapsis que transmiten información hacia otras neuronas; y finalmente las dendritas, tubos celulares que transmiten información desde otras neuronas hacia el soma.

Para realizar el símil con una neurona real, se modela una neurona teórica como una función matemática que recibe información y la transmite hacia otras neuronas de tal modo de generar salidas. Una red neuronal consiste en una serie de neuronas interconectadas, de tal forma de producir una salida (potencialmente multidimensional), aproximando alguna función real.

Diversos tipos de redes neuronales existen, pero solo se discutirán aquí las llamadas redes neuronales feed-forward, en las que la información fluye hacia un solo lado, simulando las neuronas reales. Esto porque al no haber realimentación la salida solo depende de las entradas y los pesos asignados.

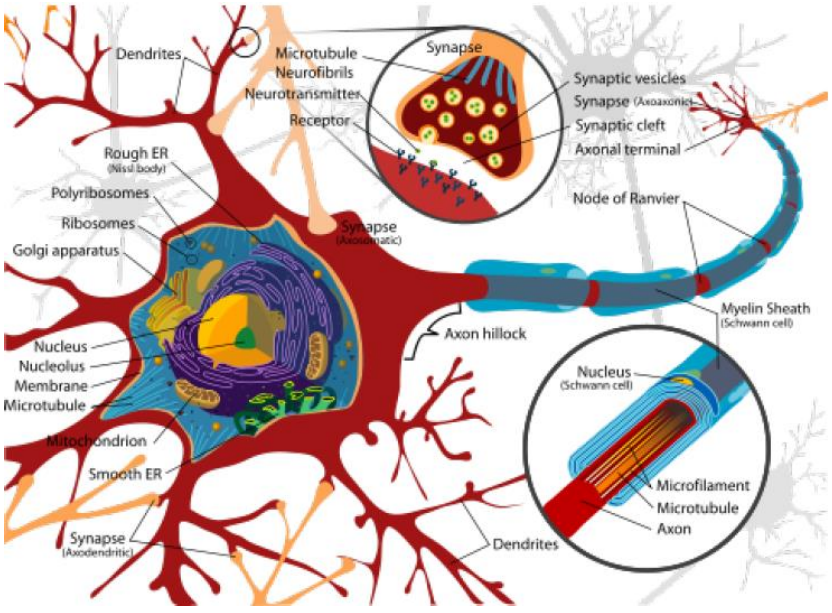

Ilustración 2. Estructura de una neurona biológica [6]

3.2.1. Estructura de una red neuronal

La estructura clásica de una red neuronal [28] corresponde a:

- **Capa(s) de entrada**

  Corresponde a la capa donde se ingresan los datos. Cada variable del modelo corresponderá a una neurona de entrada, por lo que un modelo con N variables poseerá N neuronas de entrada, cada una de ellas conectada con la primera capa oculta en base a un peso w_{nh1}.

- **Capa(s) oculta(s)**

  Corresponde al "soma" de la red neuronal. En cada una de estas capas (no necesariamente es una, sino que pueden ser varias sucesivas), las entradas de la capa anterior se mezclan utilizando alguna función matemática "de transferencia" que, a través de pesos, calcula nuevas variables y las alimenta a la siguiente capa. Funciones de transferencia clásica incluyen a la función lineal y la función sigmoidal.

- **Capa de salida**

  Corresponde a la salida de la red neuronal. Un problema con K clases tendrá asociadas K neuronas de salida, cada una de ellas aplica una función matemática a las variables que entrega la última capa oculta, ponderadas por pesos propios de la neurona de salida, y genera una variable final en el formato que se desee. Por ejemplo, si quisieran aproximar probabilidades, puede ser utilizada una función logit o softmax, mientras que si se desean aproximar algún otro número real es posible utilizar regresiones lineales o tangentes hiperbólicas.

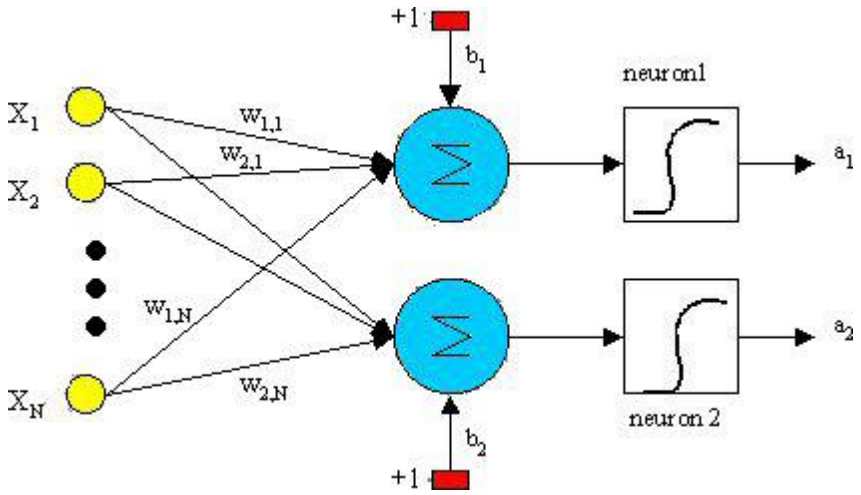

Ilustración 3. Diagrama de una red neuronal artificial [6]

3.2.2. Algoritmo backpropagation

En el punto anterior, la estructura de la red neuronal quedó definida, pero ¿cómo se ajustan los distintos pesos que se asocian a cada neurona? El algoritmo de entrenamiento de mayor fama corresponde al algoritmo backpropagation [44] que va ajustando los pesos a medida que se le presentan datos de entrenamiento, por medio de ajustar alguna función de error o "de costo", con tal de reducir el error de salida de la red.

Supongamos un conjunto de datos de entrenamiento $\{x_p, d_p\}$, donde x_p es el vector de entradas, d_p es el vector de salidas deseadas, p indica el número de ejemplo o patrón, y que la salida de la red se pueda expresar como $y = F(x; w)$, entonces el entrenamiento se realiza escogiendo el vector w que minimice la siguiente ecuación:

$$E(w) = \sum_p \left\| d_p - F(x_p; W) \right\|^2 \tag{3.1}$$

En [35] propusieron realizar el proceso de minimización usando una forma de descenso de gradiente, con la siguiente regla de actualización:

$$w_{ij}^{n+1} \leftarrow w_{ij}^n - \eta \cdot \frac{\partial E}{\partial w_{ij}} \tag{3.2}$$

Dónde:

- w_{ij} es el peso entre i y j.
- η es un parámetro que controla la velocidad de aprendizaje, esto es, la tasa de aprendizaje (Learning rate).

Rumelhart, D. y McClelland, J. observaron que las derivadas parciales se pueden calcular de manera eficiente por medio de un algoritmo conocido como retropagación de errores (error backpropagation) [35] que ha resultado ser el método estándar para realizar el entrenamiento en este tipo de redes. El algoritmo realiza la actualización, para cada peso de la red y cada patrón de entrenamiento. Al haberse realizado el proceso de entrenamiento para todos los patrones, se ha completado una época. Una vez concluido el entrenamiento, cuando el error ha disminuido aceptablemente, se puede evaluar el comportamiento de la red ante patrones no incluidos durante el entrenamiento. Esta evaluación se realiza mediante el cálculo de una medida de ajuste, como el error cuadrático medio normalizado (ECMN):

$$ECMN = \frac{\sum_{j,k}(d_{jk} - y_{jk})^2}{\sum_{j,k}(d_{jk} - \mu)^2} \tag{3.3}$$

Donde d_{jk} es el valor real de la salida j para el patrón k, y_{jk} es la predicción y μ es la media muestral de la salida real sobre todos los patrones de prueba. Las sumas son sobre todas las salidas, para todos los patrones de prueba.

Para acelerar el proceso, [35] sugirieron añadir un factor momento (momentum), α , que tiene en cuenta la dirección del incremento tomada en al iteración anterior.

Sea $\Delta w_{ij}^{n+1} = w_{ij}^{n+1} - w_{ij}^{n}$, entonces:

$$\Delta w_{ij}^{n+1} = -\eta \cdot \frac{\partial E}{\partial w_{ij}} + \alpha \cdot \Delta w_{ij}^{n} \tag{3.4}$$

3.3. Support Vector Regresion (SVR)

La Support Vector Regression (Support Vector Machine para regresiones o SVR) fue propuesta en 1996 [43].

La idea básica de SVR consiste en realizar un mapeo de los datos de entrenamiento $x \in X$, a un espacio mayor de dimensión F a través de un mapeo no lineal $\varphi : X \to F$, donde podemos realizar una regresión lineal.

3.3.1. Estructura base de SVR

Supongamos que tenemos un conjunto de datos de entrenamiento $\{(x_1, y_1),...,(x_l, y_l)\} \subset X \times \Re$, donde X corresponde al conjunto de patrones de entrada (por ejemplo $X = \Re^d$ con $d \in N$). Estos pueden ser por ejemplo las tasas de cambio para alguna moneda en los días subsiguientes junto con los indicadores econométricos correspondientes. Nuestro objetivo es encontrar una función $f(x)$ que prediga los valores y_i con una máxima desviación igual a ε para todos los datos de entrenamiento, y al mismo tiempo que sea lo más plana posible. En otras palabras, no importará el error de predicción siempre y cuando este sea menor a ε , pero no aceptará cualquier desviación más grande que ε . Esto puede ser importante si usted no desea perder más que ε de dinero cuando se trata de los tipos de cambio por ejemplo.

Para empezar se define la función lineal base para el SVR [36]:

$$f(x) = \langle w, x \rangle + b \qquad x \in X, b \in \Re \tag{3.5}$$

Dado esto, se reescribe la ecuación (3.6) de modo que cumpla con esta restricción:

$$\begin{aligned} & \min_{w,b} \frac{1}{2} \left\| w^2 \right\| \\ s.a \quad & y_i - \langle w, x_i \rangle - b \le \varepsilon \quad i = 1,...,l \\ & \langle w, x_i \rangle + b - y_i \le \varepsilon \quad i = 1,...,l \end{aligned} \tag{3.6}$$

Se toma el supuesto de que el problema de optimización convexo (3.5) es factible. A veces sin embargo este puede no ser el caso, por lo que habría que flexibilizar el valor de ε .

Análogamente a la función de pérdida suave de margen [36], que fue utilizado en las SVM por Cortés y Vapnik [42], se pueden introducir variables de holgura ξ_i, ξ_i^* para hacer frente a las restricciones infactibles del problema de optimización (3.16). Por lo tanto, se llega a la siguiente formulación:

$$\begin{aligned} & \min_{w,b} \frac{1}{2}\|w^2\| + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ s.a \quad & y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i \qquad i = 1,...,l \\ & \langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^* \qquad i = 1,...,l \\ & \xi_i, \xi_i^* \ge 0 \qquad i = 1,...,l \end{aligned} \tag{3.7}$$

La constante C>0 determina el equilibrio entre la planitud de la función f y la cantidad máxima de errores mayores que ε que serán tolerados. Esto corresponde a tratar con una función de pérdida $\varepsilon - insentive\ |\xi|_\varepsilon$:

$$|\xi|_\varepsilon := \begin{cases} 0 & si \quad |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \sim \end{cases} \tag{3.8}$$

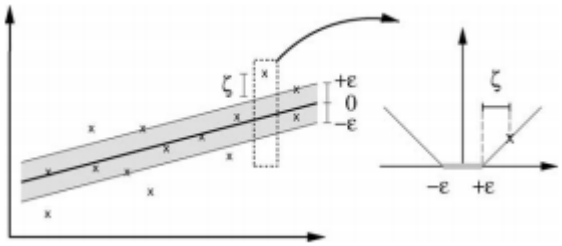

Ilustración 4. Ajuste de pérdida suave de margen en una SVM lineal [37]

La figura anterior representa gráficamente la situación. Solo los puntos fuera de la región sombreada contribuyen al costo en la función objetivo, como las desviaciones se penalizan en forma lineal. Resulta que en la mayoría de los casos el problema de optimización (3.7) se puede resolver más fácilmente en su formulación dual[2]. Además, como se verá más adelante, la formulación dual proporciona la clave para extender el SVM a funciones no lineales. Por lo tanto se utilizará un método de dualización estándar utilizando multiplicadores de Lagrange.

[2] Esto es cierto siempre y cuando la dimensionalidad de w es mucho mayor que el número de observaciones. Si este no es el caso, hay métodos especializados que pueden ofrecer un considerable ahorro computacional (Lee and Mangasarian 2001).

$$L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \tag{3.9}$$

Donde L es el Lagrangeano y $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ son los multiplicadores de Lagrange. Por lo tanto las variables duales deben cumplir con restricciones de positividad, es decir:

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0 \quad i = 1,\ldots,l \tag{3.10}$$

Si realizamos las derivadas parciales del lagrangeano con respecto a las variables primales usando las condiciones KKT tenemos:

$$\frac{\partial L(w,b,\xi_i,\xi_i^*)}{\partial b} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \tag{3.11}$$

$$\frac{\partial L(w,b,\xi_i,\xi_i^*)}{\partial w} = w - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i = 0 \tag{3.12}$$

$$\frac{\partial L(w,b,\xi_i,\xi_i^*)}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \tag{3.13}$$

$$\frac{\partial L(w,b,\xi_i,\xi_i^*)}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \tag{3.14}$$

Sustituyendo las ecuaciones (3.11), (3.12), (3.13) y (3.14) en el problema de optimización (3.9) obtenemos el siguiente problema dual:

$$\max -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l}y_i(\alpha_i - \alpha_i^*)$$

$$s.a \quad \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \tag{3.15}$$

$$\alpha_i, \alpha_i^* \in [0, C]$$

En el problema de optimización (3.14) ya se eliminaron las variables duales η_i, η_i^* porque las condiciones (3.12) y (3.13) pueden ser reformuladas de la siguiente forma:

$$\eta_i = C - \alpha_i \tag{3.16}$$

$$\eta_i^* = C - \alpha_i^* \tag{3.17}$$

Ahora podemos reescribir la ecuación (3.11) quedando de la siguiente manera:

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i \tag{3.18}$$

Esto implica que se podría redefinir la función lineal (3.15) de la siguiente manera:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \tag{3.19}$$

Esto es también llamado como la expansión del vector de soporte, es decir, w puede ser totalmente descrito como una combinación lineal de patrones de entrenamiento x_i. Esto nos muestra que la complejidad de la representación de la función por vectores soportantes es independiente de la dimensión del conjunto X al cual pertenecen los datos de entrada, y solo depende de la cantidad de vectores soportantes.

Por otra parte, el algoritmo completo se puede describir en términos de los productos punto entre los datos. Incluso a la hora de evaluar $f(x)$ no necesitamos calcular w explícitamente. Estas observaciones serán muy útiles para una formulación de extensión no lineal.

3.3.2. SVR y mapeo implícito vía función kernel

Las funciones kernel son funciones matemáticas que se emplean en las Máquinas de Soporte Vectorial. Estas funciones son las que le permiten convertir lo que sería un problema de clasificación no lineal en el espacio dimensional original, a un sencillo problema de clasificación lineal en un espacio dimensional mayor.

Para que las funciones kernel puedan ser consideradas candidatas a kernels, deben cumplir tres condiciones iniciales fundamentales; deben ser continuas, simétricas y positivas [8].

Estos son los requerimientos básicos para poder ser expresadas como un producto escalar en un espacio dimensional alto. El espacio dimensional simulado mediante las funciones kernel se define tomando a cada característica de los datos como una dimensión. Esto convierte a las entradas en un conjunto de puntos en un espacio euclidiano n-dimensional. Es mucho más fácil establecer relaciones entre los datos expresados en esta forma.

Dentro de las funciones kernel más utilizados se encuentran: el Kernel Polinomial y el de Base Radial [6].

Como se ha señalado en el punto anterior, el algoritmo de vectores soportantes depende solo del producto punto entre los patrones de entrenamiento $x_i \in X$. Por lo tanto basta con conocer una función kernel $k(x, x`) := \langle \Phi(x), \Phi(x`) \rangle$ en vez de conocer la función Φ explícitamente, lo cual permite replantear el problema de optimización de vectores de soporte (3.15):

$$\max -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$s.a \quad \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$

$$\alpha_i, \alpha_i^* \in [0, C] \tag{3.20}$$

Del mismo modo que las expresiones (3.18) y (3.19), a través de las derivadas parciales del Lagrangeano del problema a resolver se puede reescribir la restricción (3.21) y la función base (3.22) de la siguiente forma de la siguiente forma:

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\Phi(x_i) \tag{3.21}$$

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{3.22}$$

La diferencia con el caso lineal es que w ya no se da explícitamente. Hay que tener en cuenta también que en el ajuste no lineal, el problema de optimización corresponde a la búsqueda de la función más plana en el espacio característico en vez de ser la búsqueda en el espacio de los datos de entrada.

3.4. Medias Móviles (Moving Average)

El método de las medias móviles en estadística es un método utilizado para analizar un conjunto de datos en modo de puntos para crear series de promedios. Así las medias móviles son una lista de números en la cual cada uno es el promedio de un subconjunto de los datos originales.

Una serie de medias móviles puede ser calculada para cualquier serie temporal. Se usa para demanda estable, sin tendencia ni estacionalidad; suaviza las fluctuaciones de plazos cortos, resaltando así las tendencias o ciclos de plazos largos.

3.4.1. Medias Móviles Simples

Una media móvil simple es la media aritmética de los n datos anteriores, es decir, sea P_t la predicción para el momento t, y X_t el valor real obtenido en el momento t, entonces:

$$P_t = \frac{\sum_{i=1}^{n} X_{t-i}}{n} \tag{3.23}$$

En esta técnica elemental de predicción, mientras más grande sea n, mayor será la influencia de los datos antiguos. En contrapartida, si se selecciona un valor de n bajo, se tendrán en cuenta datos más recientes para la predicción.

Dependiendo del tipo de datos de serie temporal analizados podremos adaptar eficazmente la predicción a los mismos. Así, si se elige un n bajo, la predicción tendrá una alta capacidad para responder rápidamente ante fluctuaciones o variaciones significativas en los datos de un período a otro. Sin embargo, la predicción en este caso estará altamente influenciada por efectos aleatorios. Por otro lado, la elección de un n muy alto provocará que, aunque se filtre la existencia de efectos aleatorios, las predicciones presenten una adaptación lenta ante fluctuaciones significativas en los datos de períodos más recientes, pues dicha predicción estará teniendo en cuenta el valor de datos antiguos.

La simplicidad de esta técnica hace que sea objeto de críticas en lo que se refiere a su consideración equitativa de datos recientes y datos antiguos, sobre todo cuando el objeto de la predicción son variables cuya variabilidad en el corto plazo es importante para obtener una predicción eficaz. Además, en presencia de una tendencia en la serie de datos, la media móvil simple causa problemas de predicción.

3.4.2. Medias Móviles Ponderadas (Weighted Moving Average)

La media móvil ponderada es una media multiplicada por ciertos factores, que le dan determinado peso a determinados datos, es decir, sea P_t la predicción para el momento t, X_t el valor real obtenido en el momento t y W_t el peso asignado al valor real del momento t, entonces:

$$P_t = \sum_{i=1}^{n} W_{t-i} X_{t-i} \tag{3.24}$$

La media móvil ponderada desarrolla y mejora las aplicaciones de la media móvil simple. Se trata de la media aritmética de los n valores anteriores ponderados según diferentes criterios. De esta forma, se superan los inconvenientes que ofrece la técnica de media móvil simple pues, en función de las características de los datos analizados podremos decidir si darle mayor importancia a datos más antiguos o más recientes. Esta técnica será más eficiente que la media móvil simple a la hora de adaptar rápidamente el valor de la predicción a fluctuaciones en los datos recientes.

Capítulo 4

4 Simulación

Robert Shannon define la simulación como “el proceso de diseñar un modelo de un sistema real y llevar a término experiencias con él, con la finalidad de comprender el comportamiento del sistema o evaluar nuevas estrategias -dentro de los límites impuestos por un cierto criterio o un conjunto de ellos - para el funcionamiento del sistema" [39].

La idea de la simulación es realizar un experimento de una situación en un laboratorio virtual (computador) dado que es un método económico y rápido de evaluar decisiones o estrategias para el problema a evaluar. Por ejemplo, si quisiéramos evaluar el ampliar una fábrica se pueden replicar varias réplicas para cada alternativa, para así decidir cuál es la alternativa óptima de inversión y disminuir el riesgo de equivocarse, en vez de decidir a ciegas sin considerar una evaluación numérica del impacto que podría tener una decisión sobre otra.

4.1. Tipos de simulaciones

- **Simulación discreta:** Modelación de un sistema por medio de una representación en el cual los valores de sus variables de estado cambian en un conjunto numerable de instantes de tiempo (por ejemplo los sistemas de colas).
- **Simulación continua:** Modelación de un sistema por medio de una representación en el cual los valores de sus variables de estado cambian continuamente en el tiempo (por ejemplo el sistema solar).
- **Simulación combinada discreta-continua:** Modelación de un sistema por medio de una representación en el cual existen variables de estado discretas y variables de estados continuas.

  En este tipo de simulación existen 3 tipos de interacciones entre variables de distinto tipo:
  - El valor de una variable de estado continua tenga un cambio discreto gatillado por un evento discreto.
  - La relación que gobierna a una variable continua cambie en algún instante de tiempo gatillado por un evento discreto.
  - Que ocurra un evento gatillado por el valor que alcance en algún instante de tiempo una variable de estado continua.

- **Simulación determinística:** Modelación de un sistema por medio de una representación en el cual los cambios en los valores de sus variables de estado estén determinados por las condiciones iniciales del mismo (por ejemplo un sistema de ecuaciones diferenciales).
- **Simulación estocástica:** Modelación de un sistema por medio de una representación en el cual existen componentes descritos en términos probabilísticos, o donde existe incertidumbre en la entrada o en el proceso mismo del sistema (por ejemplo un sistema de colas en un banco).
- **Simulación estática:** Modelación de un sistema por medio de una representación en el cual el tiempo no juega ningún rol en el sistema (por ejemplo una simulación de Montecarlo).
- **Simulación dinámica:** Modelación de un sistema por medio de una representación en el cual el sistema evoluciona a medida que el tiempo pasa.
- **Simulación con orientación hacia los eventos:** Modelaje con un enfoque hacia los eventos, en el cual la lógica del modelo gira alrededor de los eventos que ocurren instante a instante, registrando el estado de todos los eventos, entidades, atributos y variables del modelo en todo momento.
- **Simulación con orientación hacia los procesos:** Modelaje con un enfoque de procesos, en el cual la lógica del modelo gira alrededor de los procesos que deben seguir las entidades. Es cierta forma, es un modelaje basado en un esquema de flujo grama de procesos, el cual se hace es un seguimiento a la entidad a través de la secuencia de procesos que debe seguir.

4.2. Principales paradigmas de modelamiento de simulaciones

Existen diversos tipos de paradigmas al momento de realizar simulaciones [40].

Los principales son:

- **Eventos discretos (Discrete Events o DE):** Inventado por Geoffrey Gordon en 1960s. El enfoque de su modelamiento está basado en el concepto de entidades y recursos, diagramados por gráficos de bloques que describen el flujo de entidades y el intercambio de recursos.
  Las entidades representan personas, documentos, mensajes, vehículos, etc… que viajan a través del diagrama de flujo donde pueden quedar en colas de espera, ser retrasados, procesados, usar y liberar recursos, ser separadas o combinadas, etc…
  Su enfoque es usado mayormente para decisiones tácticas y operacionales por el nivel de detalle que abarca.

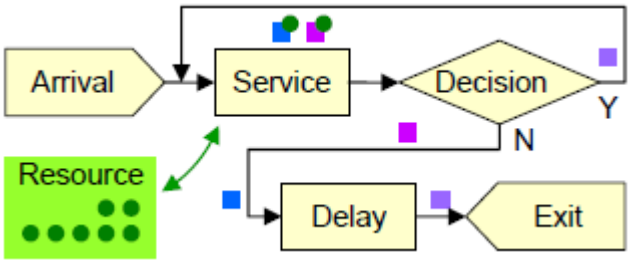

Ilustración 5. Ejemplo de un diagrama de flujo de un modelo de simulación de eventos discretos de un sistema de servicio [40]

- **Basada en Agentes (Agent Based o AB):** La idea se desarrolló como un concepto relativamente simple en la década de 1940, pero no se generalizó hasta la década de 1990.
  A este paradigma de modelamiento se le clasifica como descentralizado debido a que no está clasificado netamente ni como discreto ni como continuo. Lo que se modelo aquí es el comportamiento de cada agente y como interactúa con los demás agentes dentro de la simulación, siguiendo sus reglas de comportamiento.
  Su enfoque es utilizado tanto en decisiones estratégicas como también en decisiones tácticas y operacionales.

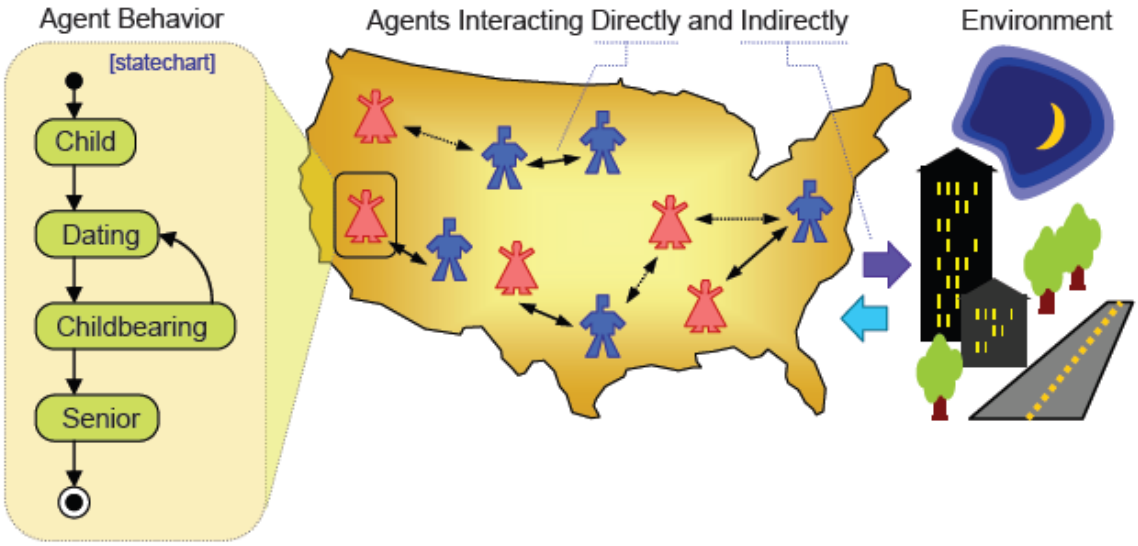

Ilustración 6. Ejemplo de una simulación basada en agentes sobre la dinámica de población de un país [40]

- **Dinámica de Sistemas (System Dynamics o SD):** Desarrollado por el ingeniero eléctrico Jay W. Forrester en la década de 1950s. El denominó esta herramienta como "el estudio de retroalimentación de información de características de la actividad industrial para mostrar como la estructura organizacional, amplificación (en políticas), y tiempos de espera (en decisiones y acciones) interactúan para influenciar el éxito de

una empresa". En SD los procesos están representados en términos de stocks, en flujos entre esos stocks e información que determinan los valores de estos flujos. SD abstrae los eventos simples y toma una vista agregada concentrándose en las políticas.
Los diagramas de flujos de este tipo de simulación se representan mediante bucles de retroalimentación.
Matemáticamente, un modelo de SD es un sistema de ecuaciones diferenciales que permiten obtener información agregada de los stocks, pero que impiden obtener información de cada ítem dentro de cada stock ya que se vuelven indistinguibles. Es por eso que este paradigma de simulación se usa para decisiones de nivel estratégico.

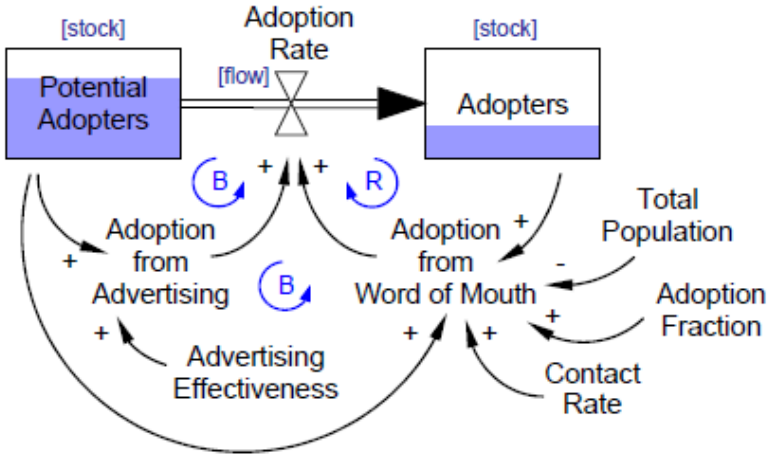

Ilustración 7. Ejemplo de una simulación de Dinámica de Sistemas sobre la adopción de un producto de parte de los clientes [40]

- **Sistemas Dinámicos (Dynamic Systems o DS):** Es el antecesor de SD. Es usado en diversas áreas de ingeniería para el diseño de procesos.
  Su modelo matemático consiste en un número de variables de estado y ecuaciones diferenciales algebraicas de varias formas de estas variables. A diferencia de SD, las variables integradas tienen un directo significado físico.
  Por el nivel de complejidad de sus técnicas se utiliza más en decisiones operacionales (ya que se usa su enfoque para el diseño en problemas de ingeniería).

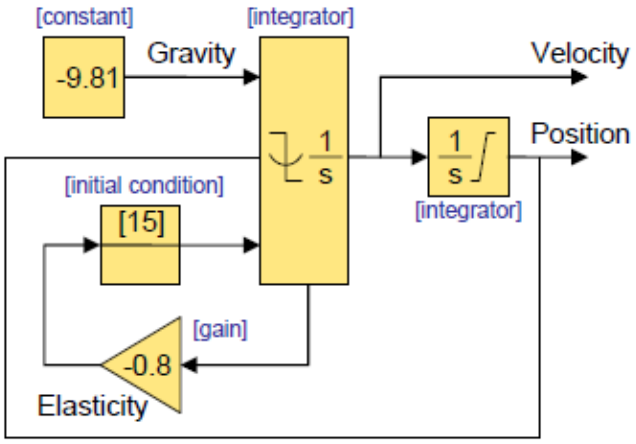

Ilustración 8. Ejemplo de una simulación de Sistemas Dinámicos sobre el rebotar de una pelota [40]

A continuación se muestra un diagrama de paradigmas de simulación versus la escala de nivel de abstracción, y además un diagrama que muestra tipos de software usados para cada paradigma de simulación.

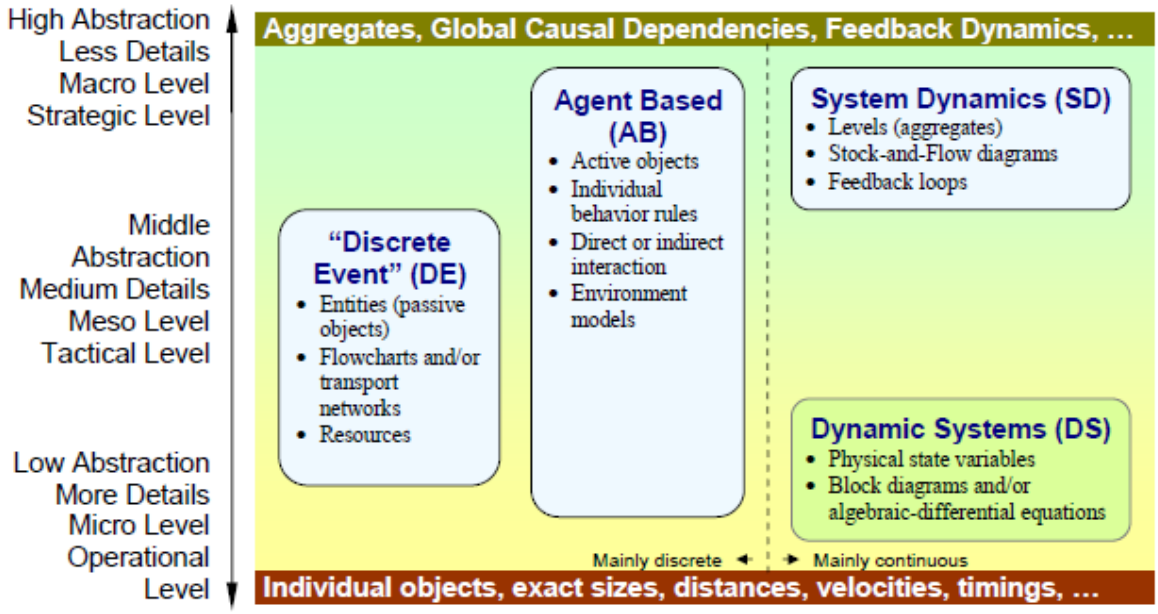

Ilustración 9. Diagrama de paradigmas de simulación con respecto a su nivel de abstracción [40]

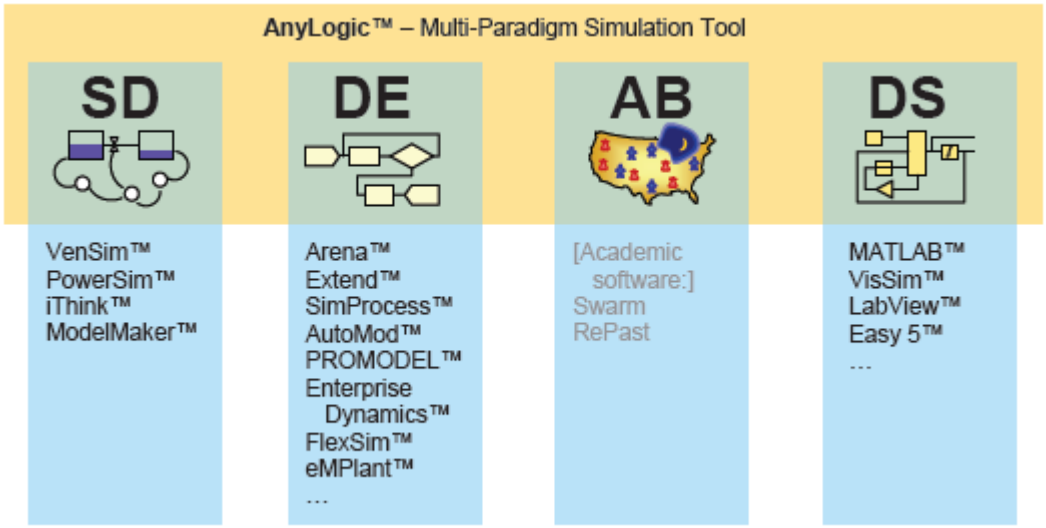

Ilustración 10. Herramientas utilizadas para cada tipo de paradigma de simulación [40]

4.3. Etapas para el desarrollo de un modelo de simulación

Las etapas para desarrollar un modelo de simulación son [16]:

- **Formular el problema:** Consiste en estudiar el contexto del problema, definir medidas de efectividad, establecer objetivos específicos a solucionar y definir el sistema que se va a modelar.
- **Recolectar los datos y definir el modelo:** Consiste en definir todas las variables que se considerarán para construir el modelo, sus relaciones lógicas y los diagramas de flujo que describan en forma completa el modelo.
- **Validación del modelo conceptual:** Consiste en validar el modelo conceptual con el cliente para ver si hay algo importante que no se esté tomando en cuenta y corregirlo, además de obtener la aprobación del cliente para continuar con el proyecto.
- **Programación del modelo:** Consiste en definir el lenguaje a utilizar para programar la simulación del modelo y obtener resultados en función de las unidades de medidas de efectividad establecidas anteriormente.
- **Validar el programa de simulación:** Consiste en verificar que el modelo simulado cumple con los requisitos de diseño para los cuales se elaboró.
  Después de esto se valida el modelo con el cliente para comprobar que los resultados de la simulación reflejan la realidad.

- **Diseño, realización y análisis de experimentos de simulación:** Consiste en evaluar posibles estrategias para solucionar el problema definido anteriormente y evaluar el impacto de estas alternativas para decidir cuál de ellas es la mejor solución.
- **Documentar y resumir los resultados de simulación:** Consiste en documentar el proyecto de simulación (tanto los aspectos técnicos como los resultados) para entregar un informe final al cliente. A veces se entrega también un manual de uso en caso de que el cliente necesite interactuar con el programa de simulación para seguir utilizándolo a futuro.

PARTE II: DESARROLLO DE LA METODOLOGÍA DE ESTUDIO

Capítulo 5

5 Modelamiento del problema

Para modelar el problema (descrito en el capítulo 1) se debe realizar un levantamiento de información del proceso a modelar, y luego se deben conocer los distintos enfoques existentes en la literatura para poder decidir cómo modelar el problema.

5.1. Levantamiento de información

En esta etapa se debe entender cómo funciona el proceso penal para poder definir responsabilidades, y saber con qué información se cuenta para realizar este estudio.

5.1.1. Descripción del proceso penal

En la ilustración 11 se aprecia un resumen del proceso penal dentro de todo el territorio nacional.

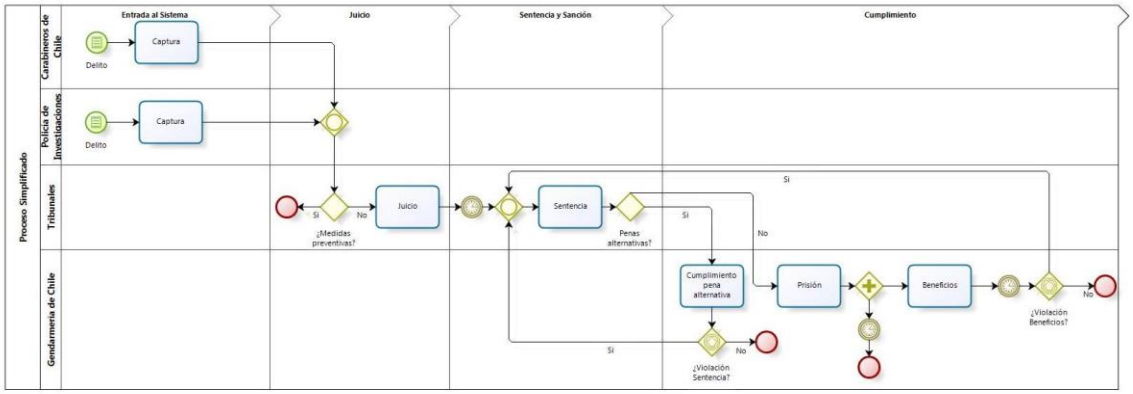

Ilustración 11. Resumen del proceso penal en base a opiniones de miembros de CEAMOS, de Gendarmería de Chile y de conocimientos propios

En función del diagrama de flujo mostrado se definen las responsabilidades de las entidades institucionales que participan de este proceso:

- Carabineros de Chile: Responsables de capturar a quienes infringen la ley[3].
- Juzgados: Responsables de llevar a cabo el juicio correspondiente y de dictar la sentencia.
- Gendarmería de Chile: Responsables de que aquellos que fueron sentenciados a cumplir condena dentro de un centro penitenciario efectivamente la cumplan.

Analizando el momento de ocurrencia de cada etapa se ve:

- Entre el delito y la captura es un tiempo aleatorio pero que mientras no sea capturado el individuo no entra a este sistema.
- El tiempo que dura el juicio hasta que se dicte la sentencia es un tiempo aleatorio que no se sabe cuánto durará porque depende de cada tipo de delito, de las evidencias, apelaciones al juicio, etc…
- El tiempo de cumplimiento de beneficios es aleatorio con un máximo determinado, ya que depende de que se cumpla la orden dictada o de lo contrario se vuelve a juicio.
- El tiempo de prisión es aleatorio porque depende de la sentencia que se dicte. Pero es un tiempo que es determinado para cada sentencia dictada, o que puede ser acortado por adquirir ciertos beneficios (como libertad condicional por ejemplo).

Dado que se busca medir la población penal de la región metropolitana se debe considerar el traslado de internos entre unidades penales de distintas regiones, ya sea por solicitud del interno o por decisión de un juez.

Otro detalle importante es que un sujeto puede estar en prisión en calidad de Imputado o en calidad de Condenado.

Definición 5.1 (Imputado): Se define Imputado a la persona sujeta a la reforma procesal penal, que está recluida en un establecimiento de Gendarmería a la que se le ha abierto proceso [17].

Definición 5.2 (Condenado): Se define Condenado a aquella persona recluida bajo la custodia de Gendarmería de Chile, la cual cumple la condena que ha dispuesto el Tribunal, en el subsistema cerrado y semiabierto [17].

[3] Se consideró solo a Carabineros de Chile y no a Policía de Investigaciones para simplificar, dado que en proporción Carabineros de Chile realizan más detenciones (449.331 [8] vs 93.077 [35] durante el año 2013).

5.1.2. Fuentes de información utilizada

La información para el desarrollo de esta tesis se obtuvo a través de 3 fuentes: Información obtenida por convenio de traspaso de información entre el Centro de Análisis y Modelamiento en Seguridad (CEAMOS) y Gendarmería de Chile, solicitud de información a Carabineros de Chile mediante ley de transparencia, y a través de preguntas a funcionarios de Carabineros de Chile apelando a su juicio de expertos.

El detalle del origen de datos para cada variable del modelo se detallará en el anexo A.

5.2. Estado del arte

Para poder estimar la cantidad de presos se han llevado a cabo diversos estudios desde distintos enfoques, como modelos econométricos y/o estadísticos y también modelos de simulación.

En el ámbito de la econometría, se puede observar un ejemplo [32] de la utilización de una serie de tiempo ARIMA que se empleó en México para predecir la población penal a futuro, no obstante, este método no busca reflejar ninguna variable explicativa para la cantidad de la población penal.

Si bien al momento de encontrar variables explicativas para la proyección de la población penal, existe una correlación de ésta con variables sociodemográficas como por ejemplo la complejidad de las ciudades (porcentaje de población urbana versus número de denuncias), el mercado laboral, la edad de los delincuentes y la educación [15, 24, 27]. Para predecir el tamaño de la población penal lo que más se suele realizar es una estimación del número de detenidos que ingresan a un centro penitenciario adulto dentro de un período determinado de tiempo y la cantidad de internos que ya se encontraban cumpliendo condena, y que continuarán realizándolo durante un período de tiempo futuro.

Un trabajo realizado por David Jacobs y Robald E. Helms [20] muestra una mezcla de predicciones mediante series de tiempo y variables explicativas al predecir la diferencia de ingresos entre períodos en función de la diferencia entre períodos de las variables explicativas sociodemográficas (personas nacidas fuera del matrimonio, variables políticas, etc.) y a la vez prediciendo el valor de las variables en función de series de tiempo.

Si bien esto mostró que la media móvil del porcentaje de personas nacidas fuera del matrimonio que tienen entre 17 y 21 años es una variable explicativa de gran impacto, muestra

que las variables políticas elegidas no eran las más relevantes al momento de elegirlas para el modelo explicativo. Sin embargo lo bueno de este estudio es que permite una metodología para poder estimar los impactos de propuestas políticas a través del tiempo en las tasas de ingresos a las cárceles.

Uno de estos estudios, realizado por Arnold Barnett [4], predice la población penal mediante procesos de Poisson, donde analiza la probabilidad de que un individuo esté cumpliendo condena a una edad determinada, lo que le permitiría analizar la probabilidad de que éste pueda ser condenado. Según lo señalado, la tasa de encarcelamiento depende de la edad de la persona, y la probabilidad de que el sujeto se encuentre libre y activo a esa edad. Por lo mismo es que estudia el período de actividad delictual. Este modelo trabaja con toda la población penal de forma agrupada, en vez de llevar a cabo una clasificación, y algo positivo que tiene es que permite anticipar cuantitativamente los cambios en la proyección de internos al ocurrir un cambio político o de actuar por parte de la policía, como también permite ajustar los parámetros dados los cambios en el tamaño y en la distribución de la edad de la población, los cambios en las sentencias impuestas en los condenados y los cambios en la probabilidad de cometer un crimen. Pero por el contrario no permite determinar cómo varían los niveles de crímenes (por tipo de delito por ejemplo) como tampoco permite medir el impacto del efecto de la incapacidad de encarcelamiento, por lo que no permite enfocar donde sería mejor realizar cambios en estas áreas.

Otros estudios se enfocan en el tiempo de reincidencia como el modelo de John MacLeod [26]. Entre estos también podemos encontrar el trabajo de Joanna R. Baker y Pamela K. Lattimore [3] en donde se muestra un modelo de flujo recursivo que separa la población en aquellos que no tienen antecedentes penales y aquellos que ya han cumplido condena al menos una vez. Este modelo pone énfasis en el período de reincidencia de aquellos que ya han cumplido condena anteriormente y en la desagregación de la información por edad y por raza. La limitación de los resultados obtenidos se sitúa principalmente en la restricción de la capacidad de presos, en tanto el nivel de hacinamiento favorecía la generación de cambios, modificándose de ésta forma los tiempos de condena, generando un impacto en el flujo recursivo.

Finalmente, resulta relevante destacar el trabajo realizado por Richard McDowall [29] quien aplicó un modelo cíclico de simulación de dinámica de sistemas, en el cual simula el "índice de hacinamiento de las cárceles del Reino Unido" en función de la población penal, la capacidad de las cárceles, la reincidencia, la rehabilitación y el punitivo Judicial. No obstante lo anterior, no considera la disuasión de los criminales, los cambios policiales, la experiencia del oficial de libertad condicional, multas y castigos alternativos, la culpabilidad o inocencia del condenado y los cambios en las tasas de reincidencia de acuerdo al tiempo del último egreso a una cárcel.

Esta tesis se diferencia de estos estudios principalmente a que se mezclará la minería de datos con la simulación para realizar predicciones y a que considera la opción de traslados de internos entre las regiones de Chile.

5.3. Factores considerados para el modelamiento del problema

Habiéndose estudiado el estado del arte, se mostrarán los siguientes factores considerados para realizar este estudio.

5.3.1. Ingresos por aprehensiones

Como se indicó en el estudio realizado por Marianov, V. [27], las cifras de aprehendidos mensuales están muy correlacionados con las cifras de ingresos mensuales a las cárceles. Pero existen dos categorías de personas que ingresan al ser aprehendidos: Aquellos que no poseen antecedentes penales (o "primerizos") y aquellos que reinciden (reincidentes).

Se sabe que para ambos casos existe un período de juicio que no se sabe a ciencia cierta cuánto durará. Lo que sí se puede comparar es la proporción de detenidos efectivos que ingresan a un centro penitenciario cerrado de la región metropolitana en un período "t" que corresponde a cada tipo de categoría de personas detenida.

A continuación se describirá cómo se modelarán las dos categorías descritas.

5.3.1.1. Reincidentes

Una de las hipótesis a evaluar en esta tesis es estudiar si aislar a los reincidentes de aquellos que no poseen antecedentes penales mejora la predicción de la población penal.

Barnett [4] modeló los ingresos de agentes con antecedentes penales de la siguiente forma:

$$\lambda(t) = \int_{h}^{\infty} \rho_t(a) FA(a) I(a) da \tag{5.1}$$

Donde:

- $\lambda(t)$ = Ingresos de reincidentes a la cárcel en el período t
- $\rho_t(a)$ = la densidad de edad específica "a" de los criminales
- $FA(a)$ = Probabilidad de que el criminal esté libre y activo a la edad de categoría "a"
- $R(a)$ = Probabilidad de que un criminal de edad "a" que está libre y activo reincida dentro de un período de tiempo
- h = Edad mínima de los delincuentes.

Dado que el modelo trabajará con la población que ya está en libertad y activa, y suponiendo que el criminal termina su carrera delictual cuando muere, entonces: sea $\rho_t^*(a) = \rho_t(a)FA(a)$, se puede calcular el valor esperado de reincidencia en el nuevo modelo de la siguiente forma:

$$E_t(\text{Re}\,incidencia) = \lambda_t = \int_h^\infty \rho_t^*(a)R(a)da \tag{5.2}$$

Tomando el supuesto de que $R(a) = R \qquad \forall a \in [c, \infty[$ quedaría entonces:

$$E_t(\text{Re}\,incidencia) = \lambda_t = \int_h^\infty \rho_t^*(a)R(a)da = \int_h^\infty \rho_t^*(a)\cdot R\cdot da = R\int_h^\infty \rho_t^*(a)da = R\cdot PA_t \tag{5.3}$$

Siendo PA_t = Población total de criminales que están libres en el período t.

A la vez se define $\lambda_{a,t} = \rho_t^*(a)\cdot R(a)$, sea A el conjunto de rangos de edad a evaluar tal que sea un conjunto discreto, entonces podemos definir:

$$E_t(\text{Re}\,incidencia) = \sum_{a\in A}\lambda_{a,t} = \sum_{a\in A}\rho_t^*(a)\cdot R(a) \tag{5.4}$$

Si fuera el caso de que $A = \{a\}$ entonces:

$$E_t(\text{Re}\,incidencia) = \sum_{a\in A}\lambda_{a,t} = \rho_t^*(a)\cdot R(a) = PA_t\cdot R \tag{5.5}$$

Este caso sería si se considerara una sola categoría en que las personas tuvieran 18 años o más.

5.3.1.2. Aprehendidos mayores de edad sin antecedentes penales

Es conocido el hecho de que cada vez más son las personas que son detenidas. Esto se puede apreciar al revisar las cifras que entrega Carabineros de Chile a través de su página de internet [7], como también en el siguiente gráfico:

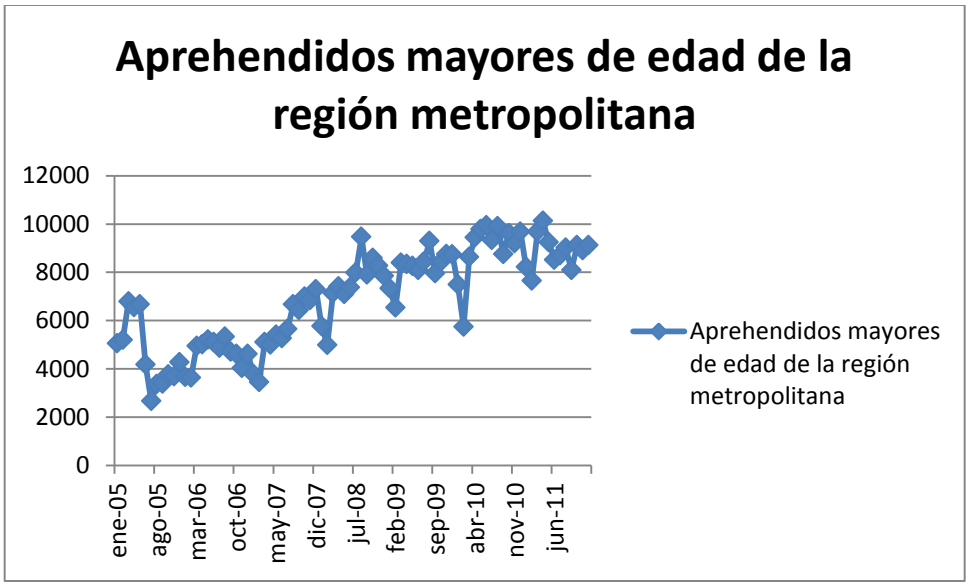

Ilustración 12. Gráfico de aprehendidos mayores de edad agrupados a nivel mensual

Es por esto que se decide modelar el ingreso de "primerizos" como una serie de tiempo no estacionaria.

Debido a la limitación de la información obtenida de la base de datos de Carabineros de Chile, no se pudo conseguir las estadísticas de aprehendidos mayores de edad sin antecedentes penales (ni siquiera por ley de transparencia). Por lo que se modela la variable de aprehendidos mayores de edad sin antecedentes penales de la siguiente forma:

- $S_{a,t}$ = Número de aprehendidos mayores de edad sin antecedentes penales dentro de la región metropolitana perteneciente al rango de edad "a" en el período t.
- $\gamma_{a,t}$ = Porcentaje de aprehendidos mayores de edad dentro de la región metropolitana que no posee antecedentes penales del rango de edad "a" en el período de tiempo t.
- $AME_{a,t}$ = Número de aprehendidos mayores de edad dentro de la región metropolitana pertenecientes al rango de edad "a" en el período de tiempo t.

Por lo tanto tendríamos:

$$S_{a,t} = AME_{a,t} \cdot \gamma_{a,t} \tag{5.6}$$

Sea S_t = Cantidad de aprehendidos sin antecedentes penales en el período t, entonces:

$$S_t = \sum_{a \in A} S_{a,t} = \sum_{a \in A} AME_{a,t} \cdot \gamma_{a,t} \tag{5.7}$$

De nuevo si asumimos el conjunto "a" como el conjunto de los mayores de edad, entonces:

$$S_t = AME_t \cdot \gamma_t \tag{5.8}$$

Debido a que S_t es una variable con tendencia al crecimiento es que se propone obtener sus valores futuros mediante una predicción de serie de tiempo usando técnicas de minería de datos.

5.3.1.3. Total ingresos por aprehensiones de Carabineros de Chile

Sea α_t la proporción de ingreso de los aprehendidos mayores de edad sin antecedentes penales a un centro penitenciario cerrado de la región metropolitana en el período t.

Sea $Ingresos_t$ = Ingresos de delincuentes a un centro penitenciario cerrado en el período "t".

Entonces:

$$Ingresos_t = S_t \cdot \alpha_t + \lambda_t \tag{5.9}$$

5.3.2. Traslados de internos entre regiones

Los internos pueden ser trasladados de un centro penitenciario a algún otro destino (tribunales, hospitales, otros centros penitenciarios, etc.) y viceversa. Además estos traslados pueden ser a otras regiones del país, y dado el enfoque regional de este estudio serán este tipo de traslados los considerados.

Por lo cual tenemos cuatro nuevas variables:

- Ingresos de Imputados por traslados desde otras regiones
- Egresos de Imputados por traslados hacia otras regiones
- Ingresos de Condenados por traslados desde otras regiones
- Egresos de Condenados por traslados hacia otras regiones

5.3.3. Escenarios y tipos de simulaciones a evaluar

Con el fin de responder unas de las hipótesis se considerarán 3 tipos de simulaciones y 2 escenarios en cada uno de ellos.

Los tipos de simulaciones a evaluar serán los siguientes:

- Simulación de dinámica de sistemas en el software Vensim, usando el valor medio de cada variable.
- Simulación de dinámica de sistemas en el software Vensim, ajustando las distribuciones de cada variable como una distribución Normal aleatoria (debido a las limitaciones de ese software).
- Simulación de eventos discretos en el software Arena, manteniendo el formato de simulación de flujo.

Y los escenarios a evaluar son los siguientes:

- Considerando la reincidencia

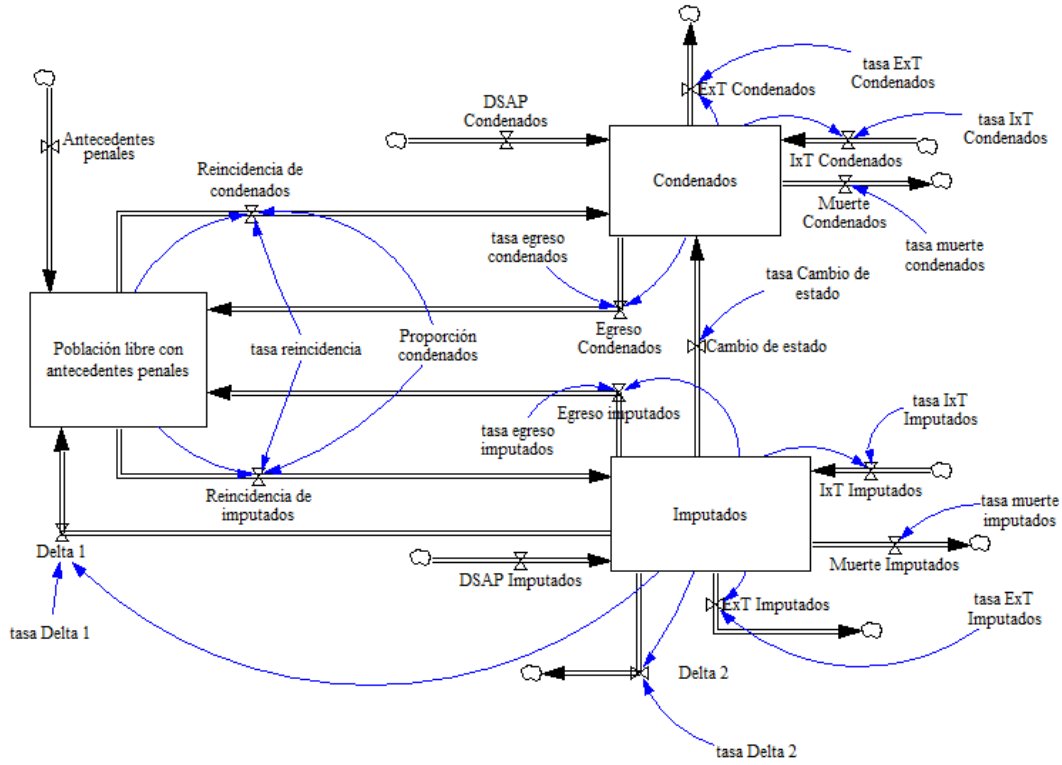

Ilustración 13. Diagrama del modelo de simulación considerando la reincidencia

- Sin considerar la reincidencia (considerar a todos los individuos de igual forma).

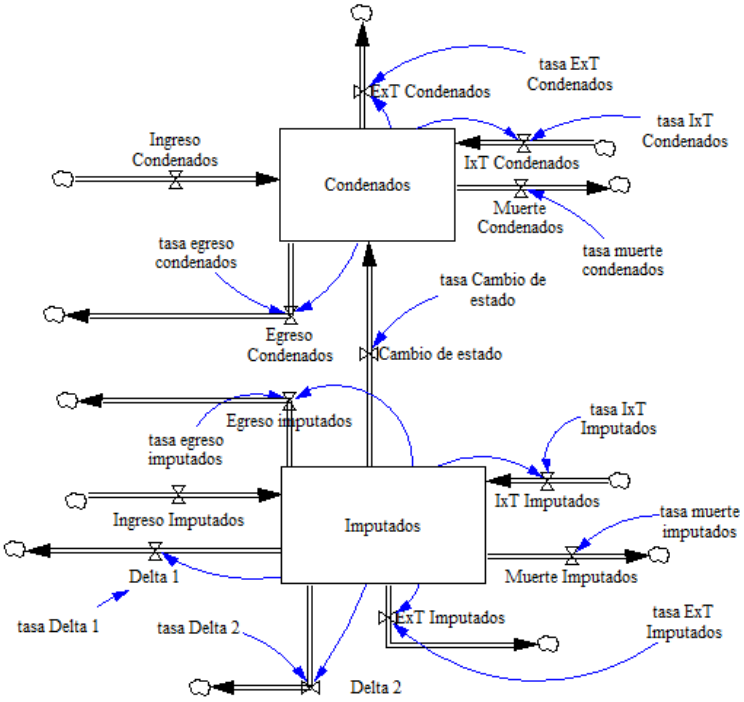

Ilustración 14. Diagrama del modelo de simulación sin considerar la reincidencia

5.4. Modelo de predicción de la población penal

A continuación se detalla el modelo de predicción de la población penal.

5.4.1. Variables aleatorias

Se definen:

- $ITC_{a,t} =$ Ingresos por traslado de condenados de rango de edad "a" a centros penitenciarios cerrados de la región metropolitana en el período "t"
- $ITI_{a,t} =$ Ingresos por traslado de imputados de rango de edad "a" a centros penitenciarios cerrados de la región metropolitana en el período "t"

- $ETC_{a,t} =$ Egresos por traslado de condenados de rango de edad “a” de centros penitenciarios cerrados de la región metropolitana en el período “t”
- $ETI_{a,t} =$ Egresos por traslado de imputados de rango de edad “a” de centros penitenciarios cerrados de la región metropolitana en el período “t”
- $CE_{a,t} =$ Cantidad de internos que cambian su estado de imputado a condenado de rango de edad “a” en centros penitenciarios cerrados de la región metropolitana en el período t
- $EI_{a,t} =$ Egresos de imputados por obtención de beneficios de rango de edad “a” de un centro penitenciario cerrado de la región metropolitana en el período “t”
- $EC_{a,t} =$ Egresos de condenados de rango de edad “a” de un centro penitenciario cerrado de la región metropolitana en el período “t”
- $MI_{a,t} =$ Muertes de imputados de rango de edad “a” en la región metropolitana en el período “t”.
- $MC_{a,t} =$ Muertes de condenados de rango de edad “a” en la región metropolitana en el período “t”.

Debido a que las variables $ITC_{a,t}, ETC_{a,t}, EC_{a,t}, ETI_{a,t}, EI_{a,t}, ITI_{a,t}$ poseen una tendencia al crecimiento, por lo que se redefinirán de forma que dependan de variables aleatorias estables y que no tengan tendencia al crecimiento o decrecimiento:

- $ITC_{a,t} = C_{t-1} \cdot PITC_{a,t}$
- $ETC_{a,t} = C_{t-1} \cdot PETC_{a,t}$
- $EC_{a,t} = C_{t-1} \cdot PEC_{a,t}$
- $ITI_{a,t} = I_{t-1} \cdot PITC_{a,t}$
- $ETI_{a,t} = I_{t-1} \cdot PETI_{a,t}$
- $EI_{a,t} = I_{t-1} \cdot PEI_{a,t}$

Donde:

- $C_{a,t} =$ Número de condenados dentro de un centro penitenciario cerrado de rango de edad “a” dentro de la región metropolitana en el período “t”.
- $PITC_{a,t} =$ Proporción de condenados de rango de edad “a” que ingresan por traslado en el período “t” a un centro penitenciario cerrado de la región metropolitana con respecto a la cantidad de condenados de rango de edad “a” dentro de un centro penitenciario cerrado de la región metropolitana en el período “t-1”.

- $PETC_{a,t} =$ Proporción de condenados de rango de edad “a” que egresan por traslado en el período “t” a un centro penitenciario cerrado de la región metropolitana con respecto a la cantidad de condenados de rango de edad “a” dentro de un centro penitenciario cerrado de la región metropolitana en el período “t-1”.
- $PEC_{a,t} =$ Proporción de condenados de rango de edad “a” que egresan en el período “t” a un centro penitenciario cerrado de la región metropolitana con respecto a la cantidad de condenados de rango de edad “a” dentro de un centro penitenciario cerrado de la región metropolitana en el período “t-1”.
- $I_{a,t} =$ Número de imputados dentro de un centro penitenciario cerrado de rango de edad “a” dentro de la región metropolitana en el período “t”.
- $PITI_{a,t} =$ Proporción de imputados de rango de edad “a” que ingresan por traslado en el período “t” a un centro penitenciario cerrado de la región metropolitana con respecto a la cantidad de imputados de rango de edad “a” dentro de un centro penitenciario cerrado de la región metropolitana en el período “t-1”.
- $PETI_{a,t} =$ Proporción de imputados de rango de edad “a” que egresan por traslado en el período “t” a un centro penitenciario cerrado de la región metropolitana con respecto a la cantidad de imputados de rango de edad “a” dentro de un centro penitenciario cerrado de la región metropolitana en el período “t-1”.
- $PEI_{a,t} =$ Proporción de imputados de rango de edad “a” que egresan en el período “t” a un centro penitenciario cerrado de la región metropolitana con respecto a la cantidad de imputados de rango de edad “a” dentro de un centro penitenciario cerrado de la región metropolitana en el período “t-1”.

5.4.2. Otros parámetros

- $CRC_{a,t} =$ Cantidad de condenados dentro de un centro penitenciario cerrado de la región metropolitana que cambia del rango de edad “a” al rango de edad “a+1” en el período “t”. (con $a \in A = \{1,2,3,...\} : a < |A|$).
- $CRI_{a,t} =$ Cantidad de imputados dentro de un centro penitenciario cerrado de la región metropolitana que cambia del rango de edad “a” al rango de edad “a+1” en el período “t”. (con $a \in A = \{1,2,3,...\} : a < |A|$).
- $CRPA_{a,t} =$ Cantidad de personas con antecedentes penales que se encuentran en libertad dentro de la región metropolitana que cambia del rango de edad “a” al rango de edad “a+1” en el período “t”. (con $a \in A = \{1,2,3,...\} : a < |A|$).

- $\delta = \delta_1 + \delta_2 =$ Proporción imputados que egresan de un centro penitenciario de la región metropolitana por algún otro motivo que no sea cumplimiento de condena (por ejemplo falta de pruebas, etc…).
- $\delta_1 =$ Proporción imputados que egresan de un centro penitenciario de la región metropolitana que si tenían antecedentes penales antes y que quedaron en libertad por algún otro motivo que no sea cumplimiento de condena (por ejemplo falta de pruebas, etc…).
- $\delta_2 =$ Proporción imputados que egresan de un centro penitenciario de la región metropolitana que no tenían antecedentes penales antes y que quedaron en libertad por algún otro motivo que no sea cumplimiento de condena (por ejemplo falta de pruebas, etc…).
- $\theta =$ Proporción de ingresos a un centro penitenciario cerrado de la región metropolitana que ingresan como calidad de condenado en un período.
- $\phi =$ Tasa de mortalidad de la región metropolitana (En el caso de la región metropolitana, la tasa de mortalidad anual promedio entre el año 2006 y 2010 fue de un 0.51%, lo cual significaría una tasa mensual de 0.0453%
- $\beta_t =$ Proporción de delincuentes sin antecedentes penales que al ser detenidos quedan con antecedentes penales pero no ingresan de inmediato a un centro penitenciario cerrado (puede quedar en un centro semiabierto o con una sentencia alternativa como arresto domiciliario por ejemplo)

5.4.3. Variable no estacionaria a predecir con serie de tiempo

La variable no estacionaria considerada en este modelo es la siguiente:

- $AME_{a,t} =$ Cantidad de Aprehendidos Mayores de Edad de rango de edad “a” en el período “t” dentro de la región metropolitana.

5.4.4. Modelo de flujo considerando la reincidencia

Habiéndose descrito las variables a ocupar, se detalla a continuación el modelo de flujo para predecir la población penal considerando la reincidencia:

- $$C_{a,t+1} = C_{a,t} + (\lambda_{a,t} + \alpha_t \cdot AME_{a,t} \cdot \gamma_{a,t}) \cdot \theta + ITC_{a,t} + CE_{a,t} - ETC_{a,t} - EC_{a,t} - MC_{a,t} - CRC_{a,t} + CRC_{a-1,t} \quad \forall a \in A : a < |A|$$
- $$C_{a,t+1} = C_{a,t} + (\lambda_{a,t} + \alpha_t \cdot AME_{a,t} \cdot \gamma_{a,t}) \cdot \theta + ITC_{a,t} + CE_{a,t} - ETC_{a,t} - EC_{a,t} - MC_{a,t} + CRC_{a-1,t} \quad a = |A|$$
- $$I_{a,t+1} = I_{a,t} \cdot (1-\delta) + (\lambda_{a,t} + \alpha_t \cdot AME_{a,t} \cdot \gamma_{a,t}) \cdot (1-\theta) + ITI_{a,t} - CE_{a,t} - ETI_{a,t} - EI_{a,t} - MI_{a,t} - CRI_{a,t} + CRI_{a-1,t} \quad \forall a \in A : a < |A|$$
- $$I_{a,t+1} = I_{a,t} \cdot (1-\delta) + (\lambda_{a,t} + \alpha_t \cdot AME_{a,t} \cdot \gamma_{a,t}) \cdot (1-\theta) + ITI_{a,t} - CE_{a,t} - ETI_{a,t} - EI_{a,t} - MI_{a,t} + CRI_{a-1,t} \quad a = |A|$$
- $$PA_{a,t+1} = PA_{a,t} \cdot (1-\phi) - \lambda_{a,t} + EC_{a,t} + EI_{a,t} + I_{a,t} \cdot \delta_1 - CRPA_{a,t} + CRPA_{a-1,t} + \beta_t \cdot AME_{a,t} \cdot \gamma_{a,t} \quad \forall a \in A : a < |A|$$
- $$PA_{a,t+1} = PA_{a,t} \cdot (1-\phi) - \lambda_{a,t} + EC_{a,t} + EI_{a,t} + I_{a,t} \cdot \delta_1 + CRPA_{a-1,t} + \beta_t \cdot AME_{a,t} \cdot \gamma_{a,t} \quad a = |A|$$
- $$C_t = \sum_{a \in A} C_{a,t}$$
- $$I_t = \sum_{a \in A} I_{a,t} \quad (5.10)$$

Para esta tesis se trabajará con una sola categoría de edad, por lo que el modelo (5.10) queda reducido al siguiente modelo de ecuaciones:

- $$C_{t+1} = C_t + (\lambda_t + \alpha_t \cdot AME_t \cdot \gamma_t) \cdot \theta + ITC_t + CE_t - ETC_t - EC_t - MC_t$$
- $$I_{t+1} = I_t \cdot (1-\delta) + (\lambda_t + \alpha_t \cdot AME_t \cdot \gamma_t) \cdot (1-\theta) + ITI_t - CE_t - ETI_t - EI_t - MI_t$$
- $$PA_{t+1} = PA_t \cdot (1-\phi) - \lambda_t + EC_t + EI_t + I_t \cdot \delta_1 + \beta_t \cdot AME_t \cdot \gamma_t \quad (5.11)$$

5.4.5. Modelo de flujo sin considerar la reincidencia

Dado que no se considera reincidencia en este modelo, todos los ingresos serán con respecto a la cantidad de detenidos en cada mes.

Para esto se define una nueva variable:

- $\eta_{a,t} =$ Porcentaje de aprehendidos mayores de edad dentro de la región metropolitana del rango de edad "a" en el período de tiempo "t" que efectivamente ingresarán a un centro penitenciario adulto.

Con esto se define el nuevo modelo de flujo de predicción de la población penal sin considerar la reincidencia.

- $C_{a,t+1} = C_{a,t} + AME_{a,t} \cdot \eta_{a,t} \cdot \theta + ITC_{a,t} + CE_{a,t} - ETC_{a,t} - EC_{a,t} - MC_{a,t} - CRC_{a,t} + CRC_{a-1,t} \quad \forall a \in A : a < |A|$
- $C_{a,t+1} = C_{a,t} + AME_{a,t} \cdot \eta_{a,t} \cdot \theta + ITC_{a,t} + CE_{a,t} - ETC_{a,t} - EC_{a,t} - MC_{a,t} + CRC_{a-1,t} \quad a = |A|$
- $I_{a,t+1} = I_{a,t} \cdot (1-\delta) + AME_{a,t} \cdot \eta_{a,t} \cdot (1-\theta) + ITI_{a,t} - CE_{a,t} - ETI_{a,t} - EI_{a,t} - MI_{a,t} - CRI_{a,t} + CRI_{a-1,t} \quad \forall a \in A : a < |A|$
- $I_{a,t+1} = I_{a,t} \cdot (1-\delta) + AME_{a,t} \cdot \eta_{a,t} \cdot (1-\theta) + ITI_{a,t} - CE_{a,t} - ETI_{a,t} - EI_{a,t} - MI_{a,t} + CRI_{a-1,t} \quad a = |A|$
- $C_t = \sum_{a \in A} C_{a,t}$
- $I_t = \sum_{a \in A} I_{a,t}$ (5.12)

Para esta tesis se trabajará con una sola categoría de edad, por lo que el modelo (5.12) queda reducido al siguiente modelo de ecuaciones:

- $C_{t+1} = C_t + AME_t \cdot \eta_t \cdot \theta + ITC_t + CE_t - ETC_t - EC_t - MC_t$
- $I_{t+1} = I_t \cdot (1-\delta) + AME_t \cdot \eta_t \cdot (1-\theta) + ITI_t - CE_t - ETI_t - EI_t - MI_t$ (5.13)

Capítulo 6

6 Cálculo de valores de los componentes del modelo predictivo y sus resultados

En este capítulo se detallará cómo se calcularon los valores de cada componente del modelo detallado en el capítulo anterior.

6.1. Cálculo del valor de los parámetros

A continuación se muestra como se realizó este paso para cada parámetro.

- **Proporción imputados que egresan de un centro penitenciario de la región metropolitana por algún otro motivo que no sea cumplimiento de condena (por ejemplo falta de pruebas, etc…) (δ):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información. Su valor corresponde a $\delta = 0{,}035$.
- **Proporción imputados que egresan de un centro penitenciario de la región metropolitana que si tenían antecedentes penales antes y que quedaron en libertad por algún otro motivo que no sea cumplimiento de condena (por ejemplo falta de pruebas, etc…) (δ_1):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información, y también calibrando manualmente junto al parámetro δ_2. Su valor corresponde a $\delta_1 = 0{,}03$.
- **Proporción imputados que egresan de un centro penitenciario de la región metropolitana que no tenían antecedentes penales antes y que quedaron en libertad por algún otro motivo que no sea cumplimiento de condena (por ejemplo falta de pruebas, etc…) (δ_2):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información, y también calibrando manualmente junto al parámetro δ_1. Su valor corresponde a $\delta_2 = 0{,}005$.

- **Proporción de ingresos a un centro penitenciario cerrado de la región metropolitana que ingresan como calidad de condenado en un período (θ):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información. Su valor corresponde al promedio del valor mensual de la variable durante el año 2011, el cual es igual a $\theta = 0{,}6$.
- **Tasa de mortalidad de la región metropolitana (ϕ):** En el caso de la región metropolitana, la tasa de mortalidad anual promedio entre el año 2006 y 2010 fue de un 0.51% (obtenida del Instituto Nacional de Estadísticas o INE), lo cual significaría una tasa mensual de 0.0453%
- **Proporción de delincuentes sin antecedentes penales que al ser detenidos quedan con antecedentes penales pero no ingresan a un centro penitenciario cerrado (β_t):** Su valor se obtuvo realizando consultas a funcionarios de Carabineros de Chile en la calle debido a que no se pudo conseguir esa información mediante ley de transparencia. Se usará un valor promedio que sea constante, el cual será $\beta_t = \beta = 0{,}4$.
- **Porcentaje de aprehendidos mayores de edad de la región metropolitana en el período de tiempo t que no posee antecedentes penales (γ_t):** Su valor se obtuvo realizando consultas a funcionarios de Carabineros de Chile en la calle debido a que no se pudo conseguir esa información mediante ley de transparencia. Se usará un valor promedio que sea constante, el cual será $\gamma_t = \gamma = 0{,}1$.
- **Proporción de ingreso de los aprehendidos mayores de edad sin antecedentes penales a un centro penitenciario cerrado de la región metropolitana en el período t (α_t):** Su valor se obtuvo realizando consultas a funcionarios de Carabineros de Chile en la calle debido a que no se pudo conseguir esa información mediante ley de transparencia. Se usará un valor promedio que sea constante, el cual será $\alpha_t = \alpha = 0{,}2$.
- **Cantidad de Condenados dentro de un centro penitenciario cerrado de la región metropolitana en el mes de Diciembre del año 2010 (C_0):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información. Su valor es de 16.272 Condenados.
- **Cantidad de Imputados dentro de un centro penitenciario cerrado de la región metropolitana en el mes de Diciembre del año 2010 (I_0):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información. Su valor es de 4.936 Imputados.

- **Cantidad de personas que están en libertad en la región metropolitana y que tienen antecedentes penales en el mes de Diciembre del año 2010 (PA_0):** Se calculó su valor usando la información de la base de datos obtenida de Gendarmería de Chile mediante el convenio de traspaso de información. Su valor es de 400.000 Condenados.
- **Probabilidad de que un criminal de edad "a" que está libre y activo reincida dentro de un período de tiempo (R):** Su valor se calculó calibrando el flujo de ingresos a las cárceles de la región metropolitana para que fuera de la misma magnitud que lo ocurrido en el año 2011. Su valor es $R = 0{,}0016$.

6.2. Cálculo de las distribuciones aleatorias de las variables estacionarias

Para cada variable aleatoria se obtuvo la distribución probabilística mediante la siguiente metodología:

- Determinar visualmente un momento en el cual se vuelva estable el registro de la variable.
- Limpiar los outliers que sean mayor o menor a 2 desviaciones estándares de la media.
- Obtener la distribución de variables aleatorias mediante la herramienta Input Analyzer (que realiza test Chi-Cuadrado y test Kolmogorov-Smirnov).

Dado el período de adaptación de la reforma penal del año 2005 (partió en las regiones de los extremos del país y se fue implementando hacia el centro hasta llegar a la región metropolitana) se decidió considerar los datos en lo posible a partir del año 2007 en adelante siempre y cuando se vean con una tendencia estable.

La obtención de la distribución aleatoria para cada variable se detalla en el anexo B.

Capítulo 7

7 Predicción de la serie de tiempo de la variable Aprehendidos Mayores de Edad

A continuación se detallará el proceso de predicción de la serie de tiempo de la variable no estacionaria Aprehendidos Mayores de edad.

7.1. Determinación de variables históricas

Dado que se cuentan con datos desde Enero del año 2005 hasta Diciembre del año 2010 se prepararon dos sets de conjuntos de posibles variables históricas explicativas.

Sea:

- x_t :La cantidad de aprehendidos mayores de edad de la región metropolitana en el mes t.
- $D_t = x_t - x_{t-1}$:La diferencia de aprehendidos mayores de edad de la región metropolitana entre el mes t y el mes t-1.
- $\tau_{i,t}$:Una variable dummy tal que vale 1 si el mes t corresponde a i (con i = {Enero, Febrero,…, Diciembre}) y cero si no.

Con esto se definen los dos sets de conjuntos de posibles variables históricas explicativas:

- $x_t = f(x_{t-12}, x_{t-24}, x_{t-36}, D_{t-12}, D_{t-24}, D_{t-36}, \tau_{Enero,t}\tau_{Febrero,t},...,\tau_{Noviembre,t},\tau_{Diciembre,t})$
- $x_t = f(x_{t-12}, x_{t-24}, D_{t-12}, D_{t-24}, \tau_{Enero,t}\tau_{Febrero,t},...,\tau_{Noviembre,t},\tau_{Diciembre,t})$

Ahora para determinar cuáles serán las variables históricas explicativas a utilizar para predecir la serie de tiempo se ordenaron de mayor a menor con respecto al peso por correlación para cada set de conjuntos usando el softwarte RapidMiner, luego se determinó un valor de corte en función de los pesos de correlación para reducir el conjunto de variables con tal de que quedaran las más significativas, y finalmente se crearon conjuntos de variables definidos por orden de pesos de correlación tal que generara un set distinto de variables quitando la variable de menor peso por correlación. Es decir: Sea A el vector de variables históricas explicativas

ordenadas de mayor a menor peso por correlación, y A_k el la variable ubicada en la posición k del vector A, entonces se define $A^n := \{A_k : 1 \leq k \leq n\}$.

Para el primer set se obtuvieron los siguientes resultados:

x_{t-12}	1
x_{t-24}	0,988904789
x_{t-36}	0,936198851
$\tau_{Febrero,t}$	0,878565675
D_{t-36}	0,483536154
D_{t-12}	0,470985285
D_{t-24}	0,469340983
$\tau_{Enero,t}$	0,408467336
$\tau_{Agosto,t}$	0,346990565
$\tau_{Diciembre,t}$	0,203740126
$\tau_{Mayo,t}$	0,135446313
$\tau_{Octubre,t}$	0,133410454
$\tau_{Junio,t}$	0,132299985
$\tau_{Noviembre,t}$	0.119344519
$\tau_{Abril,t}$	0,085845386
$\tau_{Julio,t}$	0,081033355
$\tau_{Septiembre,t}$	0,058854831
$\tau_{Marzo,t}$	0

Tabla 2. Variables históricas explicativas ordenadas de mayor a menor peso por correlación para el set de datos considerando valores de hace 36 meses atrás

Para este set de variables se determinó un valor de corte de peso de correlación igual a 0,4. Es decir, que solo se conservarán las variables que posean un peso de correlación mayor o igual a 0,4.

Por lo tanto el conjunto de variables a considerar serán las siguientes: $\{x_{t-12}, x_{t-24}, x_{t-36}, \tau_{Febrero,t}, D_{t-36}, D_{t-12}, D_{t-24}, \tau_{Enero,t}\}$

Para el segundo set de datos se obtuvieron los siguientes resultados:

x_{t-12}	1
x_{t-24}	0,911708096
$\tau_{Febrero,t}$	0,735396157
$\tau_{Enero,t}$	0,402509434
$\tau_{Agosto,t}$	0,39526333
D_{t-12}	0,384659829
D_{t-24}	0,356124395
$\tau_{Octubre,t}$	0,192205434
$\tau_{Diciembre,t}$	0,156141871
$\tau_{Noviembre,t}$	0,153036398
$\tau_{Julio,t}$	0,103048292
$\tau_{Junio,t}$	0,070290556
$\tau_{Abril,t}$	0,048452065
$\tau_{Marzo,t}$	0,044778924
$\tau_{Mayo,t}$	0,043743766
$\tau_{Septiembre,t}$	0

Tabla 3. Variables históricas explicativas ordenadas de mayor a menor peso por correlación para el set de datos considerando valores de hace 24 meses atrás

Para este set de variables se determinó un valor de corte de peso de correlación igual a 0,3. Es decir, que solo se conservarán las variables que posean un peso de correlación mayor o igual a 0,3.

Por lo tanto el conjunto de variables a considerar serán las siguientes: $\{x_{t-12}, x_{t-24}, \tau_{Febrero,t}, \tau_{Enero,t}, \tau_{Agosto,t}, D_{t-12}, D_{t-24}\}$

Finalmente se agregaron otros conjuntos de datos para evaluar con los métodos de predicción, por lo que los conjuntos definitivos a evaluar son los siguientes:

Todas
$\{x_{t-24}, \tau_{Febrero,t}\}$
$\{x_{t-12}, x_{t-24}, \tau_{Febrero,t}\}$
$\{x_{t-12}, x_{t-24}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Febrero,t}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Enero,t}, \tau_{Febrero,t}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$
$\{x_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$
$\{x_{t-12}\}$
$\{x_{t-12}, x_{t-24}, x_{t-36}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$
$\{x_{t-12}, x_{t-24}, x_{t-36}, \tau_{Febrero,t}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}\}$
$\{x_{t-12}, \tau_{Febrero,t}\}$
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$
$\{x_{t-12}, D_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$

Tabla 4. Conjuntos de variables históricas explicativas a evaluar con cada método de minería de datos para predecir la serie de tiempo de la variable Aprehendidos Mayores de edad de la región metropolitana

7.2. Predicción de la serie de tiempo

Para cada método de predicción se realizó un entrenamiento con 10 cross-validation de muestras lineales en el tiempo con los datos disponibles desde Enero del año 2005 hasta Diciembre del año 2010.

Además se midió el error de predicción en un horizonte de un año (a nivel mensual) comparando los datos predichos con los datos reales del año 2011.

Los errores que se usaron para evaluar los métodos de minería de datos usados para predecir esta serie de tiempo fueron el MAPE (Mean Absolute Percentage Error) y el MSE (Mean Squared Error).

Definición 7.1 (Mean Absolute Percentage Error o MAPE) Sea A_t el valor actual en el momento t, F_t el valor predicho para el momento t y n datos a evaluar, se define el error MAPE de la siguiente forma:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{7.1}$$

Definición 7.2 (Mean Squared Error o MSE) Sea A_t el valor actual en el momento t, F_t el valor predicho para el momento t y n datos a evaluar, se define el error MSE de la siguiente forma:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(F_t - A_t)^2 \tag{7.2}$$

A continuación se muestran las predicciones hechas con distintos métodos de predicciones.

7.2.1. Predicción de la serie de tiempo mediante el método W-LinearRegression

Se aplicó este método de predicción para cada uno de los conjuntos de variables a evaluar y en base a esto se obtuvieron las variables más significativas para la predicción, por lo que se agregaron dos nuevos conjuntos de variables: $\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembret}\}$ y $\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$.

Los resultados obtenidos con el software RapidMiner usando el método W-LinearRegression son los siguientes:

	Training		Prediction	
Variables	**MAPE**	**MSE**	**MAPE**	**MSE**
Todas	6,50%	430.535,09	6,16%	360.640,47
$\{x_{t-24}, \tau_{Febrero,t}\}$	7,63%	548.721,69	7,99%	568.300,88
$\{x_{t-12}, x_{t-24}, \tau_{Febrero,t}\}$	6,33%	399.725,86	9,58%	798.272,03
$\{x_{t-12}, x_{t-24}\}$	6,48%	456.462,69	9,29%	881.992,92
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Febrero,t}\}$	6,68%	439.129,75	9,58%	798.272,03
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,98%	429.441,26	8,92%	718.331,29
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,99%	478.396,57	6,41%	409.456,46
$\{x_{t-12}, D_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	6,42%	449.431,63	10,33%	997.881,33
$\{x_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	6,32%	440.123,13	10,39%	1.021.064,96
$\{x_{t-12}\}$	6,43%	448.576,23	8,99%	775.438,56
$\{x_{t-12}, x_{t-24}, x_{t-36}\}$	7,36%	511.921,52	13,77%	2.041.785,90
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	6,92%	512.872,92	11,07%	1.150.606,54
$\{x_{t-12}, x_{t-24}, x_{t-36}, \tau_{Febrero,t}\}$	6,46%	454.782,52	11,52%	1.278.709,46
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}\}$	6,99%	493.250,17	8,87%	797.199,46
$\{x_{t-12}, \tau_{Febrero,t}\}$	6,07%	390.662,01	9,31%	774.278,41
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,51%	395.811,43	8,82%	715.147,79
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,76%	431.163,28	8,82%	715.147,79
$\{x_{t-12}, D_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,64%	403.175,05	8,98%	736.172,68
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,34%	365.802,44	8,92%	718.331,29
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre,t}\}$	7,32%	510.993,46	6,87%	437.392,89
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	7,28%	505.905,78	7,07%	477.204,48

Tabla 5. Tabla de errores al predecir la serie de tiempo con el método W-LinearRegression

Para este método se observa que el mejor resultado se obtuvo usando todas las variables.

7.2.2. Predicción de la serie de tiempo mediante el método Redes Neuronales

Se mostrará la metodología para optimizar los parámetros del método de Redes Neuronales y después los resultados de la predicción con este método.

7.2.2.1. Optimización de parámetros

Utilizando el software RapidMiner se optimizaron los parámetros Tasa de aprendizaje y Momento para cada conjunto de variables de la siguiente forma:

a) Se realizaron iteraciones con valores de los parámetros Tasa de aprendizaje y Momento entre 0,1 y 0,9 con un delta de 0,05 (289 combinaciones).
b) En caso de que ocurriera un error en el proceso (por falta de memoria disponible para ejecutar el programa) se disminuyó e rango de la variable más alta al momento de la falla hasta el valor que permitía continuar, y así sucesivamente.
c) Se eligió la combinación que otorgó el menor MAPE.

7.2.2.2. Resultados de la predicción de la serie de tiempo mediante el método Redes Neuronales

Los resultados obtenidos fueron son los siguientes:

	Training		Prediction			
Variables	**MAPE**	**MSE**	**MAPE**	**MSE**	**Learning rate**	**Momentum**
Todas	7,55%	577.362,56	12,92%	1.939.609,21	0,35	0,1
$\{x_{t-24}, \tau_{Febrero,t}\}$	7,57%	474.636,22	16,46%	2.433.323,35	0,55	0,2
$\{x_{t-12}, x_{t-24}, \tau_{Febrero,t}\}$	6,13%	422.428,89	6,26%	352.374,39	0,25	0,25
$\{x_{t-12}, x_{t-24}\}$	6,13%	417.316,75	7,54%	565.090,53	0,35	0,1
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Febrero,t}\}$	8,18%	591.199,60	36,93%	13.201.569,38	0,1	0,1
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	7,14%	465.930,82	57,16%	21.238.527,44	0,5	0,1
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	7,20%	490.278,11	8,67%	675.760,45	0,1	0,1
$\{x_{t-12}, D_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	4,57%	279.538,92	7,22%	529.539,38	0,15	0,7
$\{x_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	5,55%	303.602,45	10,44%	1.065.998,76	0,15	0,1
$\{x_{t-12}\}$	5,70%	387.819,11	10,14%	1.087.535,28	0,6	0,3
$\{x_{t-12}, x_{t-24}, x_{t-36}\}$	6,88%	469.501,72	4,75%	272.355,18	0,1	0,7
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	5,43%	270.052,57	8,57%	790.535,81	0,55	0,25
$\{x_{t-12}, x_{t-24}, x_{t-36}, \tau_{Febrero,t}\}$	5,85%	333.061,20	8,34%	853.004,22	0,2	0,1
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}\}$	7,78%	565.020,02	8,34%	853.490,06	0,4	0,1
$\{x_{t-12}, \tau_{Febrero,t}\}$	7,37%	522.003,89	45,38%	18.528.198,27	0,1	0,5
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	7,20%	487.551,62	101,46%	81.399.755,42	0,25	0,4
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	7,39%	483.669,34	45,66%	17.229.148,99	0,1	0,1
$\{x_{t-12}, D_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,04%	340.241,59	50,52%	20.731.573,37	0,35	0,3
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,03%	386.484,46	61,61%	30.497.603,32	0,4	0,15
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre,t}\}$	6,74%	425.048,06	31,38%	8.143.626,27	0,35	0,6
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	7,04%	477.465,36	23,96%	7.727.203,88	0,1	0,15

Tabla 6. Tabla de errores al predecir la serie de tiempo con el método SVR Radial

Para este método se observa que el mejor resultado se obtuvo usando las variables $\{x_{t-12}, x_{t-24}, x_{t-36}\}$.

7.2.3. Predicción de la serie de tiempo mediante el método SVR Polinomial

Se mostrará la metodología para optimizar los parámetros del método de SVR Polinomial y después los resultados de la predicción con este método.

7.2.3.1. Optimización de parámetros

Utilizando el software RapidMiner se optimizaron los parámetros Costo y Grado para cada conjunto de variables de la siguiente forma:

a) Se realizaron iteraciones con valores de los parámetros Costo = 10^j con j=-1,0,1,…,5 y la variable Grado = k con k=1,2,…,6,7 (49 combinaciones).
b) Se guardó la combinación que otorgó el menor MAPE.
c) Luego se iteró manteniendo el valor de Grado obtenido anteriormente y se iteró el valor C (Costo) entre C/2 y 5C.
d) Se guardó la combinación que otorgó el menor MAPE.

7.2.3.2. Resultados de la predicción de la serie de tiempo mediante el método SVR Polinomial

Los resultados obtenidos fueron los siguientes:

	Training		Prediction			
Variables	**MAPE**	**MSE**	**MAPE**	**MSE**	**Grado**	**Costo**
Todas	6,950%	559.074,95	5,720%	361.591,00	1	100
$\{x_{t-24}, \tau_{Febrero,t}\}$	7,720%	559.729,01	7,930%	567.970,00	1	1.000.000
$\{x_{t-12}, x_{t-24}, \tau_{Febrero,t}\}$	6,460%	409.996,36	6,760%	417.230,77	1	50
$\{x_{t-12}, x_{t-24}\}$	6,400%	441.382,32	8,360%	650.985,38	1	100
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Febrero,t}\}$	6,540%	425.041,29	6,430%	382.037,25	1	50
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,700%	445.202,85	5,240%	505.169,77	2	100
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	5,670%	336.818,50	8,720%	772.852,48	1	50
$\{x_{t-12}, D_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	5,170%	311.136,31	9,220%	868.894,15	1	100
$\{x_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	4,740%	275.538,88	9,270%	880.438,95	1	50
$\{x_{t-12}\}$	6,650%	451.836,64	9,120%	802.680,92	1	100.000
$\{x_{t-12}, x_{t-24}, x_{t-36}\}$	5,550%	382.786,18	10,800%	1.159.177,83	1	75
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	5,340%	320.489,43	8,720%	772.852,88	1	50
$\{x_{t-12}, x_{t-24}, x_{t-36}, \tau_{Febrero,t}\}$	6,030%	408.128,81	11,240%	1.336.349,21	1	60.000
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}\}$	6,670%	463.845,77	6,530%	439.396,20	1	75
$\{x_{t-12}, \tau_{Febrero,t}\}$	6,270%	392.407,03	8,820%	695.036,11	1	1.750
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,740%	410.971,60	6,250%	395.758,82	1	50
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	7,000%	430.210,10	6,120%	361.191,21	1	50
$\{x_{t-12}, D_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,180%	351.508,92	8,970%	707.658,27	1	200.000
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,710%	414.330,09	9,090%	724.619,30	1	1.500
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre,t}\}$	8,070%	559.494,48	6,860%	459.290,00	1	100.000
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	7,750%	544.321,63	7,260%	524.083,78	1	10.000

Tabla 7. Tabla de errores al predecir la serie de tiempo con el método SVR Polinomial

Para este método se observa que el mejor resultado se obtuvo usando las variables $\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$.

7.2.4. Predicción de la serie de tiempo mediante el método SVR Radial

Se mostrará la metodología para optimizar los parámetros del método de SVR Radial y después los resultados de la predicción con este método.

7.2.4.1. Optimización de parámetros

Utilizando el software RapidMiner se optimizaron los parámetros Costo y Gamma para cada conjunto de variables de la siguiente forma:

e) Se realizaron iteraciones con valores de los parámetros Costo = 10^j con j=-1,0,1,…,5 y la variable Gamma = 2^k con k=-7,-6,…,-1,0,1,…,6,7 (105 combinaciones).
f) Se guardó la combinación que otorgó el menor MAPE.
g) Luego se iteró manteniendo el valor de Gamma obtenido anteriormente y se iteró el valor C (Costo) entre C/2 y 5C.
h) Se guardó la combinación que otorgó el menor MAPE.

7.2.4.2. Resultados de la predicción de la serie de tiempo mediante el método SVR Radial

Los resultados obtenidos fueron los siguientes:

	Training		Prediction			
Variables	**MAPE**	**MSE**	**MAPE**	**MSE**	**Gamma**	**Costo**
Todas	6,720%	451.336,62	10,020%	1.113.539,71	0,03125	20.000,00
$\{x_{t-24}, \tau_{Febrero,t}\}$	7,030%	499.804,48	8,340%	625.739,88	0,25	25.000,00
$\{x_{t-12}, x_{t-24}, \tau_{Febrero,t}\}$	6,75%	434.965,21	6,020%	345.413,63	0,03125	1.000,00
$\{x_{t-12}, x_{t-24}\}$	5,69%	384.424,63	8,86%	720.492,23	0,0625	100.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Febrero,t}\}$	6,50%	420.242,23	5,65%	286.291,78	0,015625	1.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,85%	451.374,77	5,47%	283.961,09	0,015625	1.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	4,55%	247.116,68	7,95%	622.404,97	0,03125	5.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	4,68%	245.807,22	9,08%	869.236,80	0,03125	10.000,00
$\{x_{t-12}, x_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	5,21%	297.953,12	7,02%	549.103,61	0,03125	5.000,00
$\{x_{t-12}\}$	6,55%	459.719,50	6,90%	442758,494	0,0078125	100.000,00
$\{x_{t-12}, x_{t-24}, x_{t-36}\}$	5,65%	378.013,61	9,87%	962.237,58	0,0078125	10.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, x_{t-36}, D_{t-36}, \tau_{Febrero,t}\}$	5,25%	295.709,76	9,24%	808801,196	0,015625	10.000,00
$\{x_{t-12}, x_{t-24}, x_{t-36}, \tau_{Febrero,t}\}$	4,97%	260.642,80	11,45%	1236292,727	0,03125	100.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}\}$	7,66%	540.537,11	7,48%	538169,02	0,0625	10.000,00
$\{x_{t-12}, \tau_{Febrero,t}\}$	6,57%	420.161,19	6,92%	466015,125	0,0078125	7.500,00
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,61%	395.637,47	5,44%	312699,971	0,03125	1.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,97%	418.230,74	5,89%	347796,302	0,03125	1.000,00
$\{x_{t-12}, D_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Agosto,t}\}$	6,79%	407.604,58	5,96%	341130,197	0,03125	1.000,00
$\{x_{t-12}, x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,79%	427.820,97	6,07%	352.187,61	0,03125	1.000,00
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre,t}\}$	7,03%	467.623,87	4,53%	222.519,18	0,03125	1.000,00
$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,89%	454.555,79	5,50%	315.141,31	0,125	750,00

Tabla 8. Tabla de errores al predecir la serie de tiempo con el método SVR Radial

Para este método se observa que el mejor resultado se obtuvo usando las variables $\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre,t}\}$

7.2.5. Predicción de la serie de tiempo mediante el método Medias Móviles Ponderadas Modificado

La diferencia de este método con el método de Medias Móviles Ponderadas es que los pesos asignados a las variables tienen que estar entre 0 y 1, en cambio acá solo se restringió que fueran mayores que 0.

Para utilizar este método se generó una base de datos tal que cada variable solo pudiera explicarse en función de los valores de hace 12, 24 y 36 meses, es decir: $x_t = f(x_{t-12}, x_{t-24}, x_{t-36})$.

Luego se obtuvieron los pesos asociados a las variables explicativas resolviendo el siguiente problema de programación lineal:

$$\begin{aligned} \min_{w_{t-12}, w_{t-24}, w_{t-36}} \quad & MAPE = \frac{1}{n}\sum_{t=37}^{n+36}\left|\frac{x_t - \hat{x}_t}{x_t}\right| \\ s.a \quad & \hat{x}_t = w_{t-12}\cdot x_{t-12} + w_{t-24}\cdot x_{t-24} + w_{t-36}\cdot x_{t-36} \quad \forall t = 37,\ldots,(n+36) \\ & w_{t-12}, w_{t-24}, w_{t-36} \geq 0 \quad \forall t = 37,\ldots,(n+36) \end{aligned} \tag{7.3}$$

Donde x_t es el valor real observado del número de aprehendidos mayores de edad dentro de la región metropolitana en el mes t, $\hat{x}_t$ es el valor predicho del número de aprehendidos mayores de edad dentro de la región metropolitana en el mes t, y $\{w_{t-12}, w_{t-24}, w_{t-36}\}$ los pesos asociados a las variables históricas explicativas.

Luego se evaluó el error obtenido al predecir los valores del año 2011 en función de los pesos obtenidos en el entrenamiento.

Los resultados fueron los siguientes:

w_{t-12}	w_{t-24}	w_{t-36}		**Training**	**Prediction**
0,84819137	0	0,36732191	**MAPE**	6,027%	18,694%
			MSE	349.140,41	3.439.790,68

Tabla 9. Tabla de errores al predecir la serie de tiempo con el método Medias Móviles Ponderadas Modificado

Para este método se observa que el mejor resultado se obtuvo usando las variables de hace 12 y 24 meses.

7.2.6. Resumen de resultados

A continuación se muestra una tabla con los mejores resultados obtenidos con cada método evaluado:

		Training		Prediction	
Método	**Variables**	**MAPE**	**MSE**	**MAPE**	**MSE**
W-LinearRegression	Todas	6,50%	430.535,09	6,16%	360.640,47
NeuralNet	$\{x_{t-12}, x_{t-24}, x_{t-36}\}$	6,88%	469.501,72	4,75%	272.355,18
SVgR Polinomial	$\{x_{t-12}, D_{t-12}, x_{t-24}, D_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}\}$	6,70%	445.202,85	5,24%	505.169,77
SVR Radial	$\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre t}\}$	7,03%	467.623,87	4,53%	222.519,18
W.M.A. Modificado	$\{x_{t-12}, x_{t-36}\}$	6,03%	349.140,41	18,69%	3.439.790,68

Tabla 10. Resumen de los mejores resultados de predicción de la serie de tiempo de la variable Aprehendidos Mayores de Edad de la región metropolitana obtenidos con cada método utilizado

En este caso el mejor resultado de predicción se obtuvo con el método SVR Radial, pero a la vez fue el resultado que peor MAPE de entrenamiento tuvo. Sin embargo, al ser resultados de MAPE de entrenamientos parecidos se decidió por continuar el estudio con la predicción obtenida con el método SVR Radial.

Capítulo 8

8 Resultados de las predicciones hechas con los distintos escenarios y tipos de simulaciones

En este capítulo se detallan los resultados obtenidos para cada tipo de simulación y según cada escenario a evaluar.

8.1. Resultados obtenidos con simulación de dinámica de sistemas usando el valor medio de cada variable

Los resultados fueron los siguientes:

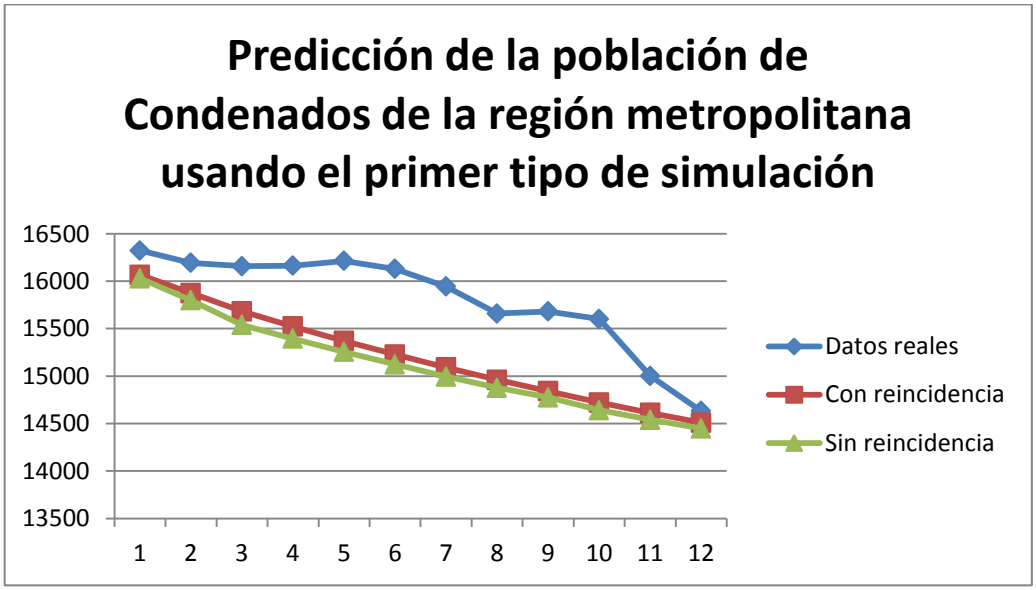

Ilustración 15. Resultados obtenidos en ambos escenarios para la población de Condenados usando simulación de dinámica de sistemas con el valor medio de cada variable

De este gráfico se puede observar que las curvas de predicción de condenados en ambos escenarios reflejan una tendencia a la baja.

Otro detalle es que el peor error absoluto de predicción para el escenario con reincidencia fue de un 5,646% para el mes 10 (Octubre del año 2011) y para el escenario sin reincidencia fue de

un 6,233% para el mes 6 (Junio del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 7% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 3,791%.
  - MSE = 432.659,847
- Escenario sin reincidencia:
  - MAPE = 4,351%
  - MSE = 554.573,154

Con estos datos se puede asumir que la predicción de Condenados usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

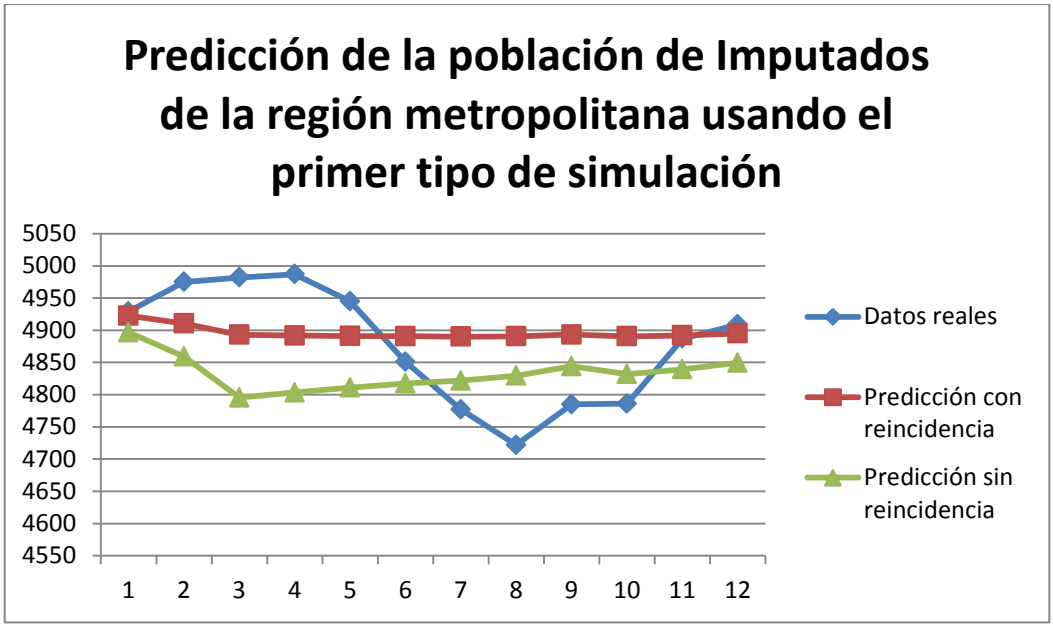

Ilustración 16. Resultados obtenidos en ambos escenarios para la población de Imputados usando simulación de dinámica de sistemas con el valor medio de cada variable

De este gráfico se puede observar que las curvas de predicción de imputados en el escenario con reincidencia reflejan una tendencia más estable que en el caso sin reincidencia.

Otro detalle es que el peor error absoluto de predicción para el escenario con reincidencia fue de un 3,569% para el mes 8 (Agosto del año 2011) y para el escenario sin reincidencia fue de un 3,745% para el mes 3 (Marzo del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 4% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 1,483%.
  - MSE = 7.476,681
- Escenario sin reincidencia:
  - MAPE = 1,783%
  - MSE = 10.576,7355

Con estos datos se puede asumir que la predicción de Imputados usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

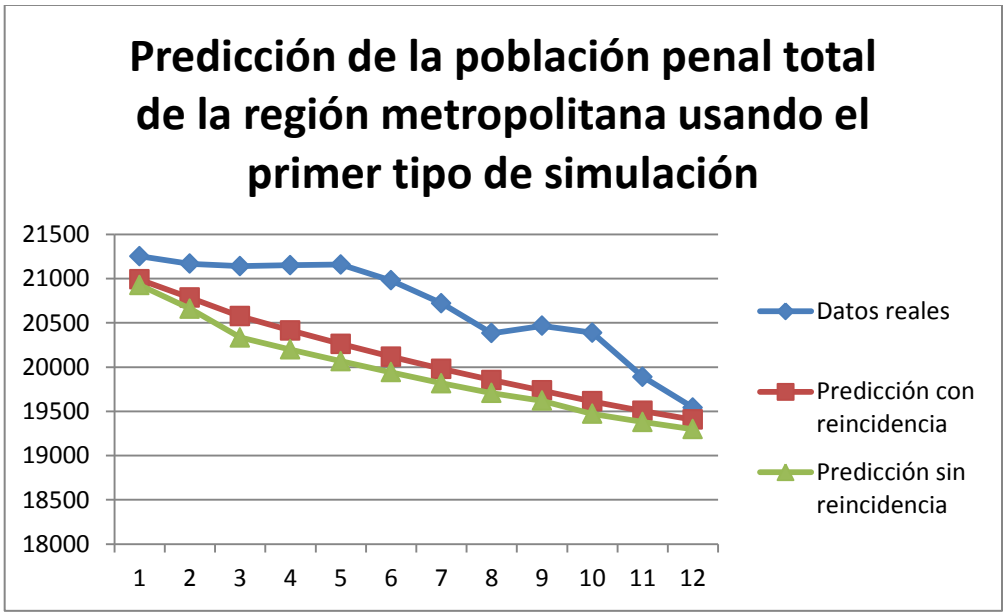

Ilustración 17. Resultados obtenidos en ambos escenarios para la población penal total usando simulación de dinámica de sistemas con el valor medio de cada variable

De este gráfico se puede observar que las curvas de predicción de la población penal total en ambos escenarios reflejan una tendencia a la baja.

Para poder medir el error de predicción de la población penal se definió un error ponderado con respecto al tipo de población (condenado o imputado) ya que lo es más importante predecir cada tipo de población por separado que la población agrupada. Esto porque no se pueden mezclar Imputados con Condenados dentro de un centro penitenciario y además porque si el error de predicción de cada tipo de población es en sentido contrario (es decir, que

en una predicción se prediga que habrán más internos de lo real y en la otra se prediga menos internos de lo real) el error agrupado se pueda compensar y anular).

Definición 8.1 (Error agrupado para la población penal total) Se define el error agrupado para la población penal total de la siguiente forma:

$$Error_{e,t} = \frac{C_t}{C_t + I_t} \cdot e(C) + \frac{I_t}{C_t + I_t} \cdot e(I) \tag{8.1}$$

Donde e corresponde al tipo de error a medir (MAPE, MSE, etc.), t corresponde al instante de tiempo en el cual se están midiendo los errores, C_t es la cantidad de Condenados en la región metropolitana para el instante de tiempo t, y I_t es la cantidad de Imputados en la región metropolitana para el instante de tiempo t.

Con esto se puede decir que el peor error absoluto de predicción para el escenario con reincidencia fue de un 4,834% para el mes 10 (Octubre del año 2011) y para el escenario sin reincidencia fue de un 5,159% para el mes 5 (Mayo del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 6% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 2,751%
  - MSE = 251.450,88
- Escenario sin reincidencia:
  - MAPE = 3,332%
  - MSE = 336.455,69

Con estos datos se puede asumir que la predicción de la población penal total usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

A continuación podemos ver una tabla resumen con los errores MAPE y MSE para cada tipo de población y escenario obtenidos dentro de un horizonte de un año:

		MAPE	MSE
Condenados	Con reincidencia	3,791%	432.659,85
	Sin reincidencia	4,351%	554.573,15
Imputados	Con reincidencia	1,483%	7.476,68
	Sin reincidencia	1,783%	10.576,74
Población total	Con reincidencia	2,751%	251.450,88
	Sin reincidencia	3,332%	336.455,69

Tabla 11. Tabla de errores MAPE y MSE obtenidos con la simulación de dinámica de sistemas con el valor medio de cada variable

Esta tabla muestra que para cada tipo de población y tipo de error, se obtuvo una mejor predicción considerando el escenario con reincidencia para el horizonte total de un año.

Sin embargo falta poder medir hasta que horizonte de tiempo es confiable este resultado.

Para esto se calculó el MAPE y el MSE para cada tipo de población y escenario con respecto a la cantidad de meses considerados como horizonte de tiempo.

Se define un error MAPE límite de 4% máximo para que sea un resultado confiable.

Primero se mostrará la tabla para el escenario con reincidencia:

	Condenados		Imputados		Población total	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	1,568%	65536	0,127%	39,44	1,234%	50.346,02
2	1,767%	83.457,28	0,713%	2.103,72	1,377%	57.341,61
3	2,162%	131.449,39	1,070%	4.036,29	1,553%	72.035,66
4	2,614%	201.467,61	1,279%	5.287,27	1,739%	92.829,75
5	3,130%	303.202,74	1,241%	4.808,92	1,929%	120.957,02
6	3,541%	388.269,62	1,171%	4.272,63	2,106%	150.711,15
7	3,800%	436.892,95	1,342%	5.486,40	2,267%	177.386,25
8	3,882%	443.199,29	1,620%	8.350,89	2,404%	198.019,94
9	4,045%	472.112,44	1,692%	8.729,36	2,525%	216.436,88
10	4,205%	502.499,68	1,741%	8.950,12	2,635%	233.457,28
11	4,059%	470.623,90	1,593%	8.138,94	2,709%	244.687,08
12	3,791%	432.659,85	1,483%	7.476,68	2,751%	251.450,88

Tabla 12. Valores MAPE y MSE obtenidos para cada tipo de población con respecto al horizonte de tiempo de predicción definido en meses durante un año con la simulación de dinámica de sistemas usando los valores medios para cada variable y considerando la reincidencia

Se denota con color amarillo los meses para los cuales un error MAPE superó el valor 4%, los cuales en este caso fueron los meses 9, 10 y 11 (correspondientes a los meses Septiembre, Octubre y Noviembre del año 2011) en la predicción de la población de condenados, pero que nunca superan el 4,3%.

Sin embargo el MAPE agrupado para la población penal no superó el 2,76% lo cual es un resultado muy bueno que permite decir que el resultado es confiable dentro de un horizonte de un año.

Ahora se mostrarán los mismos resultados para el escenario sin reincidencia

	Condenados		Imputados		Población total	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	1,801%	86436	0,655%	1.042,00	1,535%	66.631,41
2	2,115%	120.521,12	1,491%	7.222,35	1,752%	80.262,27
3	2,689%	208.563,43	2,242%	16.418,93	2,029%	107.936,00
4	3,207%	304.455,13	2,602%	20.732,26	2,288%	140.341,68
5	3,747%	427.078,59	2,622%	20.165,23	2,527%	178.669,38
6	4,162%	524.370,35	2,300%	16.990,39	2,728%	216.733,46
7	4,418%	578.280,32	2,105%	14.848,25	2,893%	249.826,86
8	4,491%	582.729,31	2,127%	14.438,63	3,024%	274.982,32
9	4,634%	609.105,10	2,028%	13.222,43	3,135%	296.627,21
10	4,787%	640.700,51	1,921%	12.111,70	3,233%	316.278,67
11	4,632%	601.901,03	1,835%	11.216,52	3,298%	329.049,86
12	4,351%	554.573,15	1,783%	10.576,74	3,332%	336.455,69

Tabla 13. Valores MAPE y MSE obtenidos para cada tipo de población con respecto al horizonte de tiempo de predicción definido en meses durante un año con la simulación de dinámica de sistemas usando los valores medios para cada variable y sin considerar la reincidencia

Aquí se observa que para la predicción para la población de condenados sobrepasa el 4% de MAPE a partir del mes 6 (Junio del año 2011) en adelante, con valores de MAPE mayores a los del escenario con reincidencia.

Si bien el MAPE agrupado para la población total penal nunca superó el 4%, los resultados fueron peores que al considerar la reincidencia.

Con todo esto se puede concluir definitivamente que para este tipo de simulación las predicciones para cada tipo de población fueron mejores usando el escenario con reincidencia, los cuales también fueron resultados confiables durante todo el año.

8.2. Resultados obtenidos con simulación de dinámica de sistemas ajustando las distribuciones de cada variable como una distribución Normal aleatoria

Los resultados fueron los siguientes:

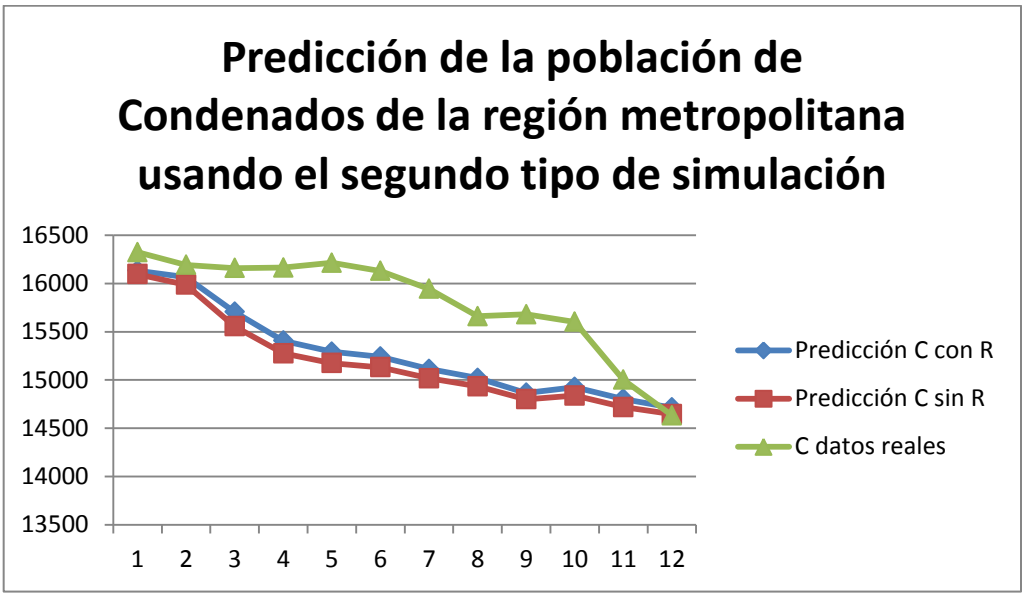

Ilustración 18. Resultados obtenidos en ambos escenarios para la población de Condenados usando simulación de dinámica de sistemas ajustando las distribuciones de cada variable como distribuciones Normales

De este gráfico se puede observar que las curvas de predicción de condenados en ambos escenarios reflejan una tendencia a la baja.

Otro detalle es que el peor error absoluto de predicción para el escenario con reincidencia fue de un 5,679% para el mes 5 (Mayo del año 2011) y para el escenario sin reincidencia fue de un 6,410% para el mes 5 (Mayo del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 7% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 3,446%
  - MSE = 394.604,416
- Escenario sin reincidencia:
  - MAPE = 3,951%
  - MSE = 511.897,024

Con estos datos se puede asumir que la predicción de Condenados usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

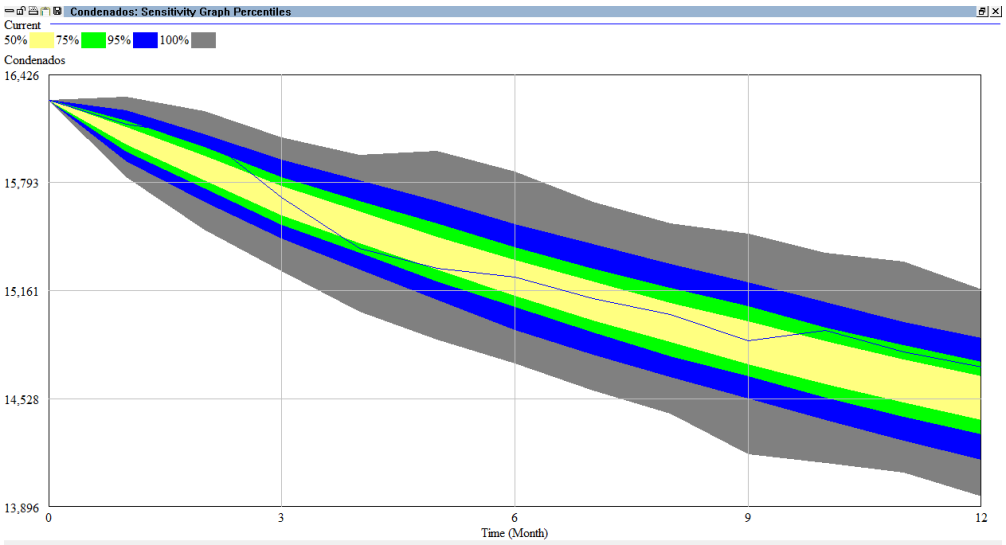

Ilustración 19. Gráfico de sensibilidad de la predicción de la población de Condenados usando simulación de dinámica de sistemas ajustando las distribuciones de las variables como distribuciones Normales, en el escenario con reincidencia

Aquí se observa un gráfico de sensibilidad para la predicción de la población de condenados de la región metropolitana con el escenario de reincidencia. En el cual el color amarillo muestra que con 50% de probabilidad la población de condenados estará en ese rango, el verde suma el rango extra para el cual sería el valor de la población de condenados sería con un 75% de probabilidad, y así sucesivamente el color azul sería para un 95% de probabilidad y el color gris para un 100% de los casos simulados.

De este gráfico se puede deducir que el rango de valores se mantiene acotado a medida que avanza el horizonte de predicción.

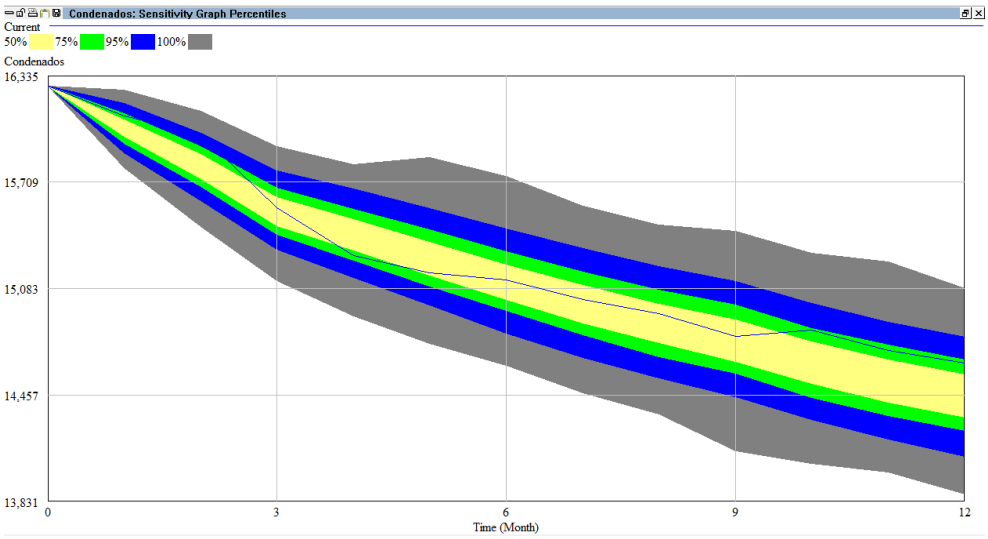

Ilustración 20. Gráfico de sensibilidad de la predicción de la población de Condenados de la región metropolitana para el año 2011 usando simulación de dinámica de sistemas ajustando las distribuciones de las variables como distribuciones Normales, en el escenario sin reincidencia

Este gráfico de sensibilidad para la predicción de condenados sin considerar la reincidencia sigue la misma estructura de colores que el gráfico anterior.

De este gráfico se observa que el rango de valores también se mantiene acotado a medida que avanza el horizonte de predicción.

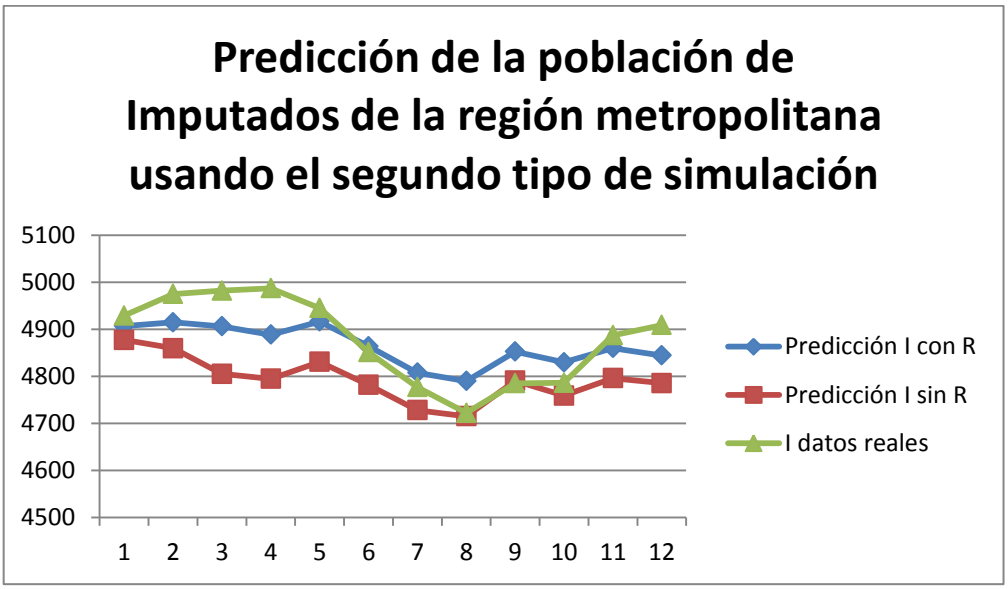

Ilustración 21. Resultados obtenidos en ambos escenarios para la población de Imputados usando simulación de dinámica de sistemas ajustando las distribuciones de cada variable como distribuciones Normales

De este gráfico se puede observar que las curvas de predicción de imputados en el escenario con reincidencia reflejan una tendencia más estable que en el caso sin reincidencia, y que este último escenario predice en su mayoría valores menores que los reales.

Otro detalle es que el peor error absoluto de predicción para el escenario con reincidencia fue de un 1,976% para el mes 4 (Abril del año 2011) y para el escenario sin reincidencia fue de un 3,856% para el mes 4 (Abril del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 4% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 1,022%.
  - MSE = 3.122
- Escenario sin reincidencia:
  - MAPE = 1,730%
  - MSE = 10.743,218

Con estos datos se puede asumir que la predicción de Imputados usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

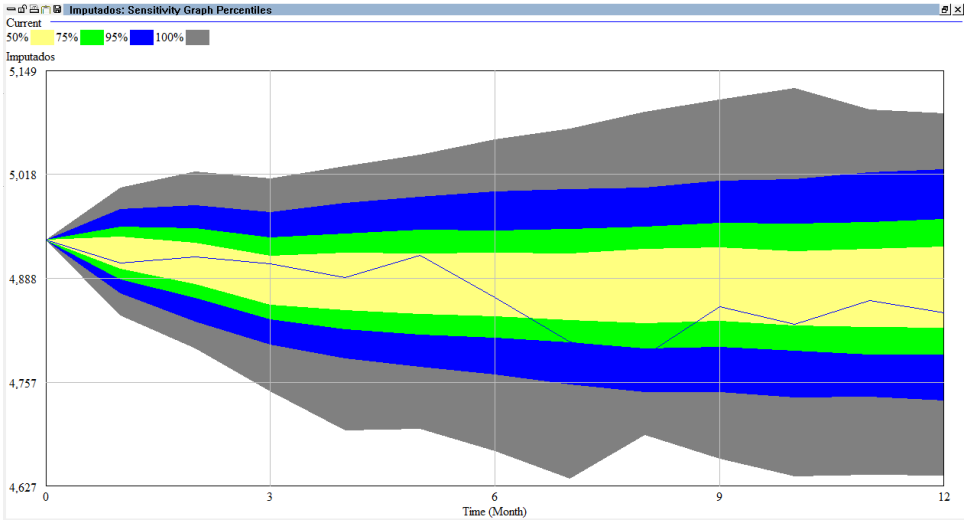

Ilustración 22 Gráfico de sensibilidad de la predicción de la población de Imputados de la región metropolitana para el año 2011 usando simulación de dinámica de sistemas ajustando las distribuciones de las variables como distribuciones Normales, en el escenario con reincidencia

Aquí se tiene el gráfico de sensibilidad de la predicción de Imputados usando el escenario con reincidencia. Si bien se aprecia que el intervalo de confianza aumenta a medida que aumenta el horizonte de tiempo, este se mantiene de todas formas acotado ya que para un 95% de confianza se aprecia visualmente que el rango mide aproximadamente 250 internos (lo cual es

aproximadamente un 5% de la magnitud de la población de imputados de la región metropolitana).

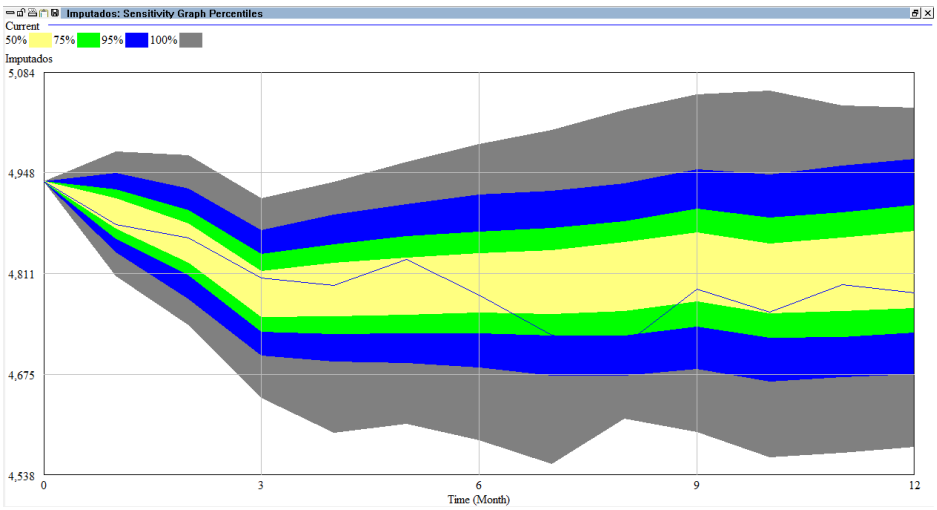

Ilustración 23. Gráfico de sensibilidad de la predicción de la población de Imputados de la región metropolitana para el año 2011 usando simulación de dinámica de sistemas ajustando las distribuciones de las variables como distribuciones Normales, en el escenario sin considerar la reincidencia

Aquí se tiene el gráfico de sensibilidad de la predicción de Imputados usando el escenario sin reincidencia. Si bien se aprecia que el intervalo de confianza aumenta a medida que aumenta el horizonte de tiempo, este se mantiene de todas formas acotado ya que para un 95% de confianza se aprecia visualmente que el rango mide aproximadamente 300 internos (lo cual es aproximadamente un 6% de la magnitud de la población de imputados de la región metropolitana). Pero igual es menos acotado que en el escenario con reincidencia.

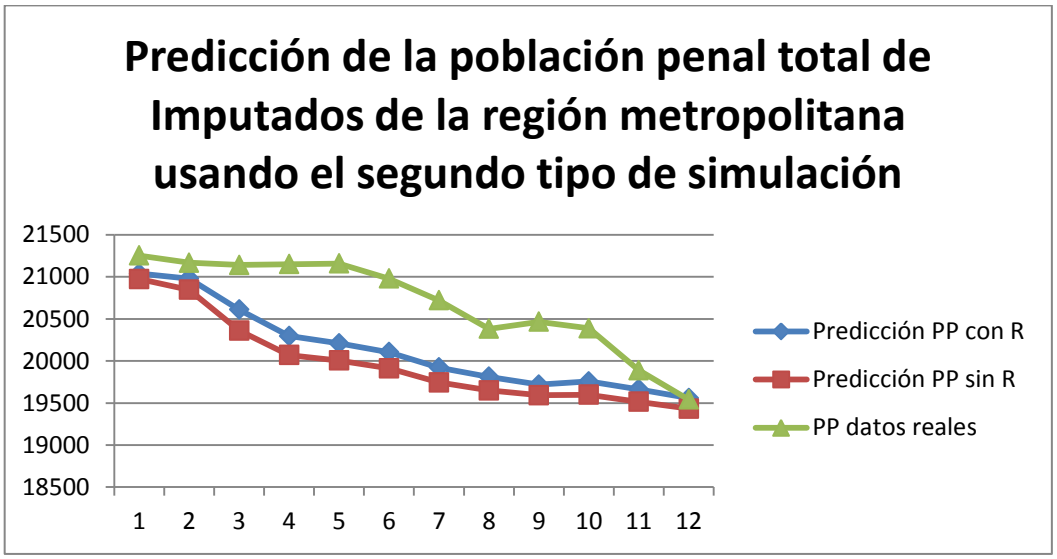

Ilustración 24. Resultados obtenidos en ambos escenarios para la población penal total usando simulación de dinámica de sistemas ajustando las distribuciones de cada variable como distribuciones Normales

De este gráfico se puede observar que las curvas de predicción de la población penal total en ambos escenarios reflejan una tendencia a la baja.

Con esto se puede decir que el peor error absoluto de predicción para el escenario con reincidencia fue de un 4,486% para el mes 5 (Mayo del año 2011) y para el escenario sin reincidencia fue de un 5,453% para el mes 5 (Mayo del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 6% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 2,463%
  - MSE = 240.417,44
- Escenario sin reincidencia:
  - MAPE = 3,112%
  - MSE = 324.585,30

Con estos datos se puede asumir que la predicción de la población penal total usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

		MAPE	MSE
Condenados	Con reincidencia	3,446%	394.604,42
	Sin reincidencia	3,951%	511.897,02
Imputados	Con reincidencia	1,022%	3.122,00
	Sin reincidencia	1,730%	10.743,22
Población total	Con reincidencia	2,463%	240.417,44
	Sin reincidencia	3,112%	324.585,30

Tabla 14. Tabla de errores MAPE y MSE obtenidos con la simulación de dinámica de sistemas ajustando las distribuciones de cada variable como distribuciones Normales

Esta tabla muestra que para cada tipo de población y tipo de error, se obtuvo una mejor predicción considerando el escenario con reincidencia para el horizonte total de un año.

Sin embargo falta poder medir hasta que horizonte de tiempo es confiable este resultado.

Para esto se calculó el MAPE y el MSE para cada tipo de población y escenario con respecto a la cantidad de meses considerados como horizonte de tiempo.

Primero se mostrará la tabla para el escenario con reincidencia:

	Condenados		Imputados		Población total	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	1,175%	36787,24	0,451%	494,62	1,007%	28.370,25
2	0,985%	26.662,60	0,833%	2.073,20	0,978%	24.626,87
3	1,593%	86.480,40	1,063%	3.299,37	1,141%	38.710,68
4	2,365%	208.009,02	1,291%	4.902,06	1,384%	69.063,08
5	3,028%	335.981,75	1,148%	4.083,41	1,625%	106.933,44
6	3,441%	411.497,60	1,001%	3.430,28	1,834%	141.968,57
7	3,694%	451.221,40	0,949%	3.072,61	2,009%	171.388,14
8	3,742%	445.954,74	1,010%	3.266,36	2,146%	192.888,99
9	3,904%	470.188,88	1,054%	3.406,54	2,268%	211.574,00
10	3,947%	468.894,64	1,040%	3.255,72	2,367%	226.375,38
11	3,709%	429.875,18	0,995%	3.025,97	2,429%	235.340,57
12	3,446%	394.604,42	1,022%	3.122,00	2,463%	240.417,44

Tabla 15. Valores MAPE y MSE obtenidos para cada tipo de población con respecto al horizonte de tiempo de predicción definido en meses durante un año con la simulación de dinámica de sistemas ajustando las distribuciones de cada variable como distribución Normal y considerando la reincidencia

Vemos que en ningún instante el valor MAPE superó el 4% para ningún tipo de población, por lo que se considera un resultado confiable para el horizonte de un año.

Ahora se observará el escenario sin reincidencia:

	Condenados		Imputados		Población total	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	1,393%	51710,76	1,049%	2.672,89	1,313%	40.337,89
2	1,334%	47.155,86	1,684%	7.996,18	1,365%	39.145,13
3	2,132%	152.359,00	2,306%	15.752,56	1,634%	66.152,39
4	2,972%	311.316,46	2,694%	21.060,20	1,952%	110.334,19
5	3,659%	465.082,07	2,618%	19.469,30	2,245%	160.455,23
6	4,081%	553.469,27	2,420%	17.024,83	2,487%	205.284,77
7	4,328%	597.137,03	2,221%	14.938,66	2,680%	242.089,45
8	4,366%	588.216,16	1,962%	13.077,29	2,822%	268.699,63
9	4,504%	608.903,25	1,757%	11.627,72	2,937%	290.983,98
10	4,543%	606.260,35	1,638%	10.539,37	3,030%	308.527,31
11	4,301%	558.416,30	1,658%	10.332,41	3,086%	319.001,61
12	3,951%	511.897,02	1,730%	10.743,22	3,112%	324.585,30

Tabla 16. Valores MAPE y MSE obtenidos para cada tipo de población con respecto al horizonte de tiempo de predicción definido en meses durante un año con la simulación de dinámica de sistemas ajustando las distribuciones de cada variable como distribución Normal y sin considerar la reincidencia

Aquí se observa que para la predicción para la población de condenados sobrepasa el 4% de MAPE a partir del mes 6 (Junio del año 2011) hasta el mes 11 (Noviembre del año 2011), con valores de MAPE mayores a los del escenario con reincidencia.

Si bien el MAPE agrupado para la población total penal nunca superó el 4%, los resultados fueron peores que al considerar la reincidencia.

Con todo esto se puede concluir definitivamente que para este tipo de simulación las predicciones para cada tipo de población fueron mejores usando el escenario con reincidencia, los cuales también fueron resultados confiables durante todo el año.

8.3. Resultados obtenidos con simulación de eventos discretos

Los resultados fueron los siguientes:

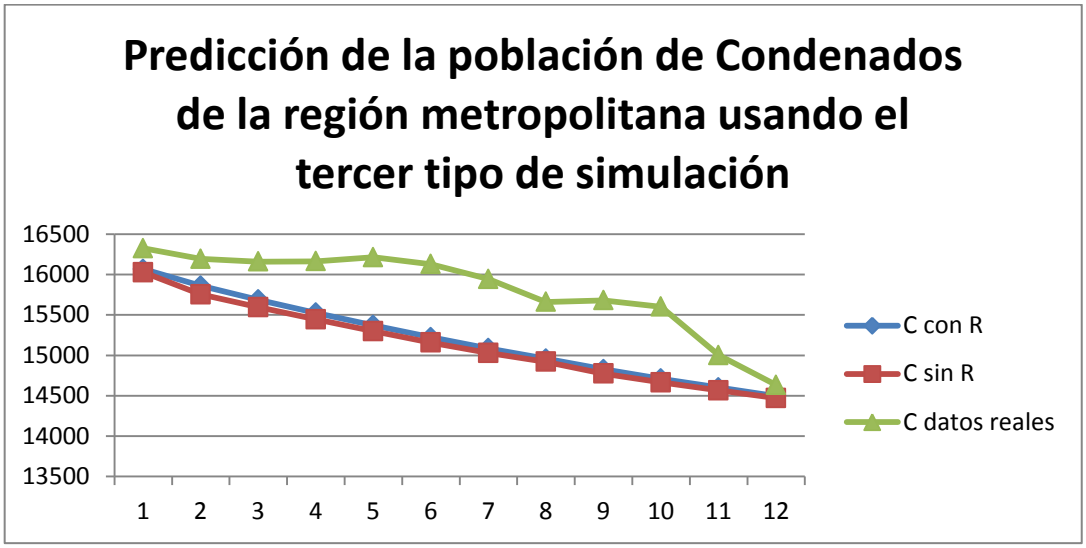

Ilustración 25. Resultados obtenidos en ambos escenarios para la población de Condenados usando simulación de eventos discretos

De este gráfico se puede observar que las curvas de predicción de condenados en ambos escenarios reflejan una tendencia a la baja.

Otro detalle es que el peor error absoluto de predicción para el escenario con reincidencia fue de un 5,718% para el mes 10 (Octubre del año 2011) y para el escenario sin reincidencia fue de un 6,024% para el mes 6 (Junio del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 7% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 3,829%
  - MSE = 439.359,402
- Escenario sin reincidencia:
  - MAPE = 4,202%
  - MSE = 517.014,259

Con estos datos se puede asumir que la predicción de Condenados usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

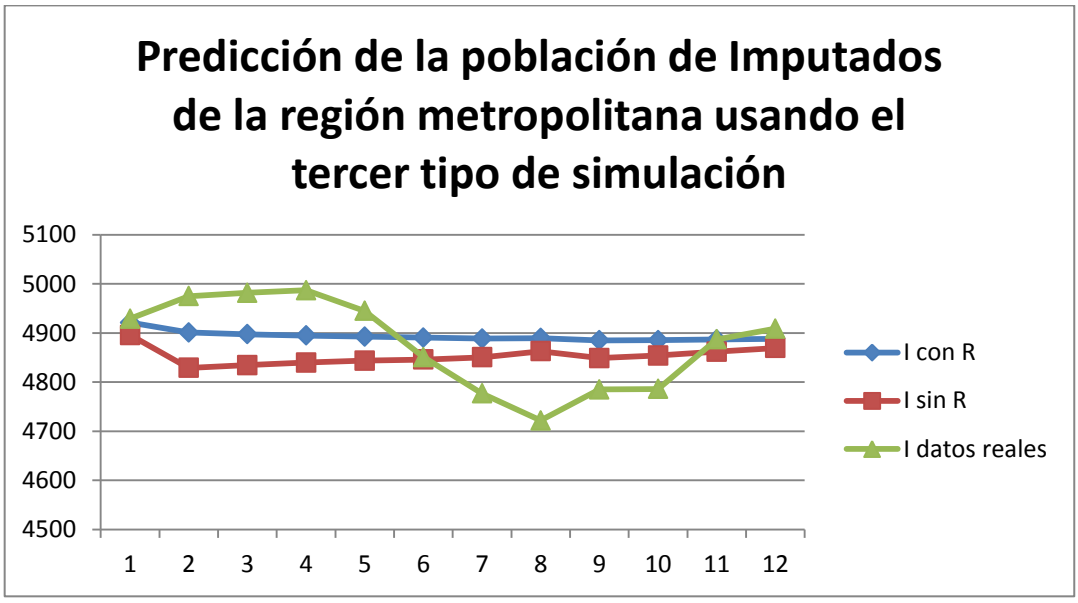

Ilustración 26. Resultados obtenidos en ambos escenarios para la población de Imputados usando simulación de eventos discretos

De este gráfico se puede observar que las curvas de predicción de imputados en el escenario con reincidencia reflejan una tendencia más estable que en el caso sin reincidencia.

Otro detalle es que el peor error absoluto de predicción para el escenario con reincidencia fue de un 3,549% para el mes 8 (Agosto del año 2011) y para el escenario sin reincidencia fue de un 2,988% para el mes 8 (Agosto del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 4% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 1,462%.
  - MSE = 7.195,73
- Escenario sin reincidencia:
  - MAPE = 1,689%
  - MSE = 9.360,98

Con estos datos se puede asumir que la predicción de Imputados usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

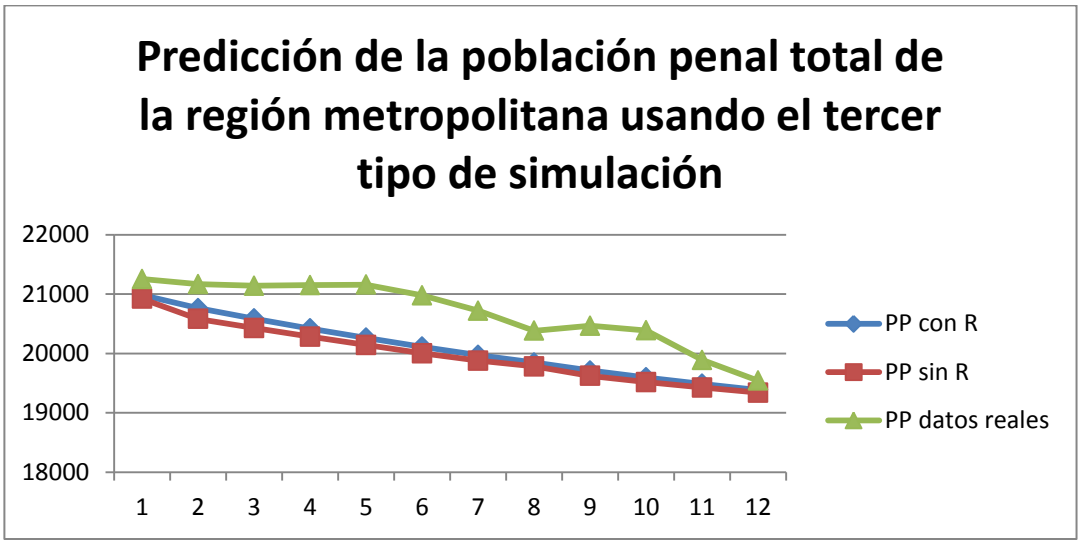

Ilustración 27. Resultados obtenidos en ambos escenarios para la población penal total usando simulación de eventos discretos

De este gráfico se puede observar que las curvas de predicción de la población penal total en ambos escenarios reflejan una tendencia a la baja.

Con esto se puede decir que el peor error absoluto de predicción para el escenario con reincidencia fue de un 4,864% para el mes 10 (Octubre del año 2011) y para el escenario sin reincidencia fue de un 4,933% para el mes 5 (Mayo del año 2011). Por lo que en ninguno de los dos escenarios tiene un error mayor a un 5% para ningún mes en particular.

Finalmente para cada escenario evaluado se tuvieron los siguientes valores de MAPE y MSE:

- Escenario con reincidencia:
  - MAPE = 2,769%
  - MSE = 253.851,84
- Escenario sin reincidencia:
  - MAPE = 3,239%
  - MSE = 314.253,86

Con estos datos se puede asumir que la predicción de la población penal total usando el escenario con reincidencia fue mejor que usando el escenario sin reincidencia.

		MAPE	MSE
Condenados	Con reincidencia	3,829%	439.359,40
	Sin reincidencia	4,202%	517.014,26
Imputados	Con reincidencia	1,462%	7.195,73
	Sin reincidencia	1,689%	9.360,98
Población total	Con reincidencia	2,769%	253.851,84
	Sin reincidencia	3,239%	314.253,86

Tabla 17. Tabla de errores MAPE y MSE obtenidos con la simulación de eventos discretos

Esta tabla muestra que para cada tipo de población y tipo de error, se obtuvo una mejor predicción considerando el escenario con reincidencia para el horizonte total de un año.

Sin embargo falta poder medir hasta que horizonte de tiempo es confiable este resultado.

Para esto se calculó el MAPE y el MSE para cada tipo de población y escenario con respecto a la cantidad de meses considerados como horizonte de tiempo.

Primero se mostrará la tabla para el escenario con reincidencia:

	Condenados		Imputados		Población total	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	1,581%	66.622,46	0,159%	61,56	1,251%	51.185,64
2	1,815%	88.374,19	0,822%	2.761,88	1,417%	59.719,42
3	2,179%	132.500,81	1,113%	4.213,37	1,587%	73.902,66
4	2,621%	201.107,49	1,296%	5.274,54	1,767%	94.160,45
5	3,137%	303.103,69	1,248%	4.766,52	1,953%	122.004,42
6	3,550%	389.323,41	1,176%	4.233,51	2,128%	151.717,49
7	3,812%	438.999,92	1,342%	5.415,43	2,287%	178.478,10
8	3,896%	445.802,85	1,618%	8.248,17	2,422%	199.222,38
9	4,069%	477.389,12	1,671%	8.444,80	2,543%	217.947,54
10	4,234%	509.240,71	1,712%	8.592,07	2,653%	235.324,33
11	4,092%	477.617,94	1,557%	7.810,98	2,727%	246.856,66
12	3,829%	439.359,40	1,462%	7.195,73	2,769%	253.851,84

Tabla 18. Valores MAPE y MSE obtenidos para cada tipo de población con respecto al horizonte de tiempo de predicción definido en meses durante un año con la simulación de eventos discretos considerando la reincidencia

Se denota con color amarillo los meses para los cuales un error MAPE superó el valor 4%, los cuales en este caso fueron los meses 9, 10 y 11 (correspondientes a los meses Septiembre, Octubre y Noviembre del año 2011) en la predicción de la población de condenados, pero que nunca superan el 4,3%.

Sin embargo el MAPE agrupado para la población penal no superó el 2,77% lo cual es un resultado muy bueno que permite decir que el resultado es confiable dentro de un horizonte de un año.

Ahora se mostrará los mismos resultados para el escenario sin reincidencia:

	Condenados		Imputados		Población total	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
1	1,814%	87706,76	0,674%	1.102,74	1,550%	67.621,54
2	2,265%	140.614,60	1,804%	11.213,64	1,853%	88.911,87
3	2,674%	199.797,54	2,190%	14.726,17	2,089%	111.336,04
4	3,119%	279.467,08	2,380%	16.464,12	2,303%	137.866,04
5	3,625%	391.196,07	2,313%	15.215,71	2,506%	170.958,22
6	4,024%	483.318,25	1,945%	12.684,00	2,679%	204.881,53
7	4,269%	533.734,74	1,887%	11.646,80	2,828%	234.666,04
8	4,326%	535.599,42	2,025%	12.678,61	2,948%	257.139,25
9	4,489%	567.560,77	1,948%	11.722,29	3,053%	277.190,93
10	4,641%	598.698,48	1,896%	11.015,72	3,148%	295.546,08
11	4,483%	561.617,81	1,770%	10.070,76	3,209%	307.414,17
12	4,202%	517.014,26	1,689%	9.360,98	3,239%	314.253,86

Tabla 19. Valores MAPE y MSE obtenidos para cada tipo de población con respecto al horizonte de tiempo de predicción definido en meses durante un año con la simulación de eventos discretos sin considerar la reincidencia

Aquí se observa que para la predicción para la población de condenados sobrepasa el 4% de MAPE a partir del mes 6 (Junio del año 2011en adelante, con valores de MAPE mayores a los del escenario con reincidencia.

Si bien el MAPE agrupado para la población total penal nunca superó el 4%, los resultados fueron peores que al considerar la reincidencia.

Con todo esto se puede concluir definitivamente que para este tipo de simulación las predicciones para cada tipo de población fueron mejores usando el escenario con reincidencia, los cuales también fueron resultados confiables durante todo el año.

8.4. Resumen de resultados de los distintos tipos de simulaciones

		Tipo de simulación					
		S.D.		S.D aleatorio		Discreta	
Tipo de población	Escenario	MAPE	MSE	MAPE	MSE	MAPE	MSE
Condenados	Con reincidencia	3,791%	432.659,85	3,446%	394.604,42	3,829%	439.359,40
	Sin reincidencia	4,351%	554.573,15	3,951%	511.897,02	4,202%	517.014,26
Imputados	Con reincidencia	1,483%	7.476,68	1,022%	3.122,00	1,462%	7.195,73
	Sin reincidencia	1,783%	10.576,74	1,730%	10.743,22	1,689%	9.360,98
Población total	Con reincidencia	2,751%	251.450,88	2,463%	240.417,44	2,769%	253.851,84
	Sin reincidencia	3,332%	336.455,69	3,112%	324.585,30	3,239%	314.253,86

Tabla 20. Resumen de los errores de predicción MAPE y MSE para los distintos tipos de población, dependiendo del tipo de simulación utilizado y del escenario considerado (S.D. = Dinámica de sistemas)

La tabla anterior nos muestra los errores MAPE y MSE obtenidos al predecir la población de Condenados, Imputados y de la población penal total, dependiendo del tipo de simulación utilizado y del escenario considerado.

Se destaca con color amarillo los mejores resultados obtenidos al predecir cada tipo de población. Y en este caso resultó que se obtuvieron los mejores resultados utilizando simulación de dinámica de sistemas con distribuciones aleatorias.

Y otro aspecto importante es que para cada tipo de simulación y para cada tipo de población, la predicción fue mejor considerando la reincidencia que sin considerarla.

Ahora si bien la simulación de dinámica de sistemas obtuvo la mejor predicción, las limitaciones de la licencia del software Vensim utilizada para el desarrollo de esta tesis da un resultado más confiable para el resultado gráfico de sensibilidad de la simulación, la predicción numérica obtenida a través de él fue generada por una semilla en particular. Por lo tanto se compararán el primer y el tercer tipo de simulación con respecto al resultado de predicción exacto y se guardará el resultado gráfico de sensibilidad de la simulación obtenida por la simulación del segundo tipo.

		Tipo de simulación			
		S.D.		Discreta	
Tipo de población	Escenario	MAPE	MSE	MAPE	MSE
Condenados	Con reincidencia	3,791%	432.659,85	3,829%	439.359,40
	Sin reincidencia	4,351%	554.573,15	4,202%	517.014,26
Imputados	Con reincidencia	1,483%	7.476,68	1,462%	7.195,73
	Sin reincidencia	1,783%	10.576,74	1,689%	9.360,98
Población total	Con reincidencia	2,751%	251.450,88	2,769%	253.851,84
	Sin reincidencia	3,332%	336.455,69	3,239%	314.253,86

Tabla 21. Resumen de los errores de predicción MAPE y MSE para los distintos tipos de población, dependiendo del tipo de simulación utilizado y del escenario considerado. Sin considerar la simulación del segundo tipo (S.D. = Dinámica de sistemas)

En esta tabla se muestra que la simulación de dinámica de sistemas entrega predicciones con menores MAPE para la población de Condenados y de la población penal total, la simulación de eventos discretos entrega una mejor predicción para la población de Imputados.

Pero si consideramos que la población de Condenados es casi un 75% de la población penal total y que se obtiene un menor MAPE para la predicción de la población penal total con la simulación de dinámica de sistemas que con la simulación de eventos discretos, se puede concluir que la simulación de dinámica de sistemas entrega mejores resultados de predicción que la simulación de eventos discretos para este problema en particular. Lo cual concuerda con que este problema es más estratégico que operacional al predecirse la población penal de la región metropolitana en vez de para cada cárcel de la región, por lo que desde un principio se podía intuir que convenía utilizar este tipo de simulación.

Sin embargo vemos que la mejora los resultados al complejizar el modelo fueron más bien marginales. Y si a esto le sumamos que en este estudio se consideró el caso simplificado con un solo rango etario (mayores de edad) y a que el dato más valioso para la predicción de un período es el del mes anterior (ya que la población penal existente no se crea de la nada) es que se puede concluir que el seguir complejizando más el modelo no aportará mejoras significativas a la calidad de las predicciones.

8.5. Predicción de la población penal a 36 meses

Dado que se determinó el mejor tipo de simulación, ahora falta predecir la proyección de la población penal de la región metropolitana en un horizonte de 3 años.

Para esto hay que predecir la serie de tiempo de los Aprehendidos Mayores de edad de la región metropolitana durante 3 años para tener los inputs necesarios para el modelo de simulación.

8.5.1. Predicción de la serie de tiempo de Aprehendidos Mayores de edad de la región metropolitana en un horizonte de 3 años

Antes de predecir esta serie de tiempo se debe analizar la estructura de la predicción obtenida para esta serie en el capítulo 6.

Las variables históricas explicativas utilizadas en el modelo SVR Radial fueron $\{x_{t-24}, \tau_{Enero,t}, \tau_{Febrero,t}, \tau_{Septiembre,t}\}$, por lo que el método elegido permite predecir a un horizonte de 2 años con datos reales. Por lo que la predicción para cada uno de los 3 años (2011, 2012 y 2013) se hará de la siguiente forma:

- Predicción para el año 2011: Se realizará con los datos del año 2009.
- Predicción para el año 2012: Se realizará con los datos del año 2010.
- Predicción para el año 2013: Se realizará con los datos predichos para el año 2011.

8.5.2. Resultados de la predicción a 3 años

Con los inputs obtenidos en el paso anterior se obtuvieron las siguientes predicciones:

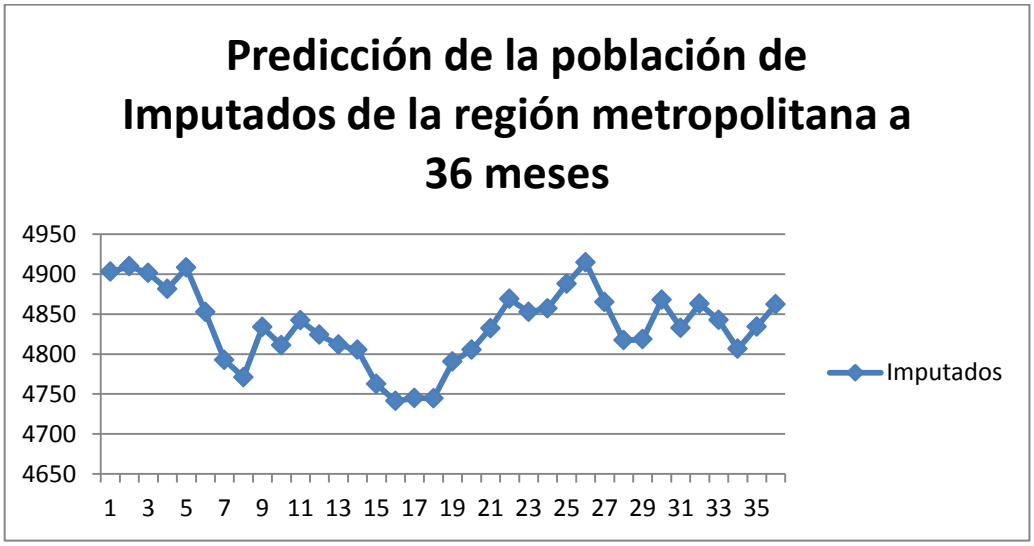

Ilustración 28. Gráfico de predicción a 36 meses de la población de imputados de la región metropolitana

Se observa que la magnitud de la población de Imputados se mantiene dentro del mismo rango.

Ahora si observamos el gráfico de sensibilidad se observa lo siguiente:

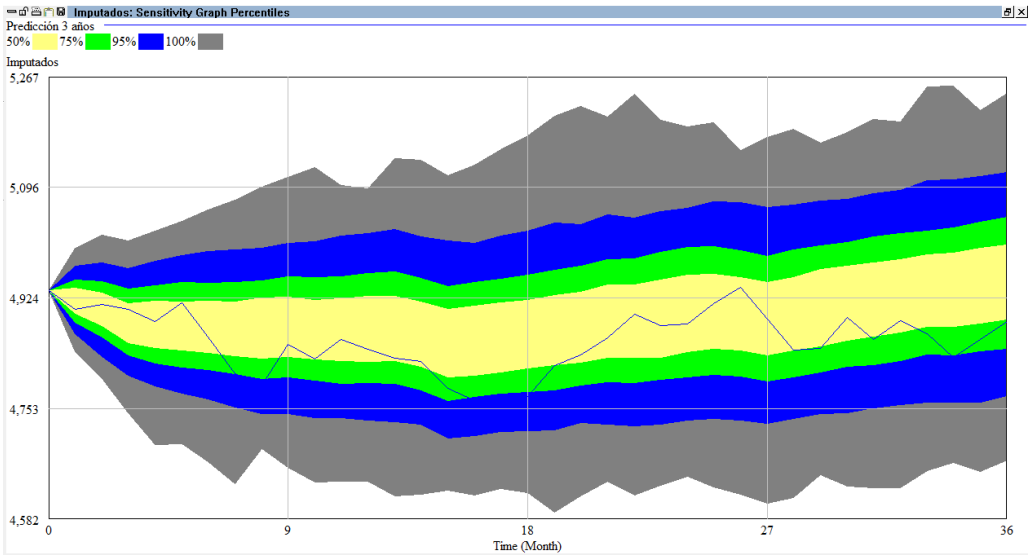

Ilustración 29. Gráfico de análisis de sensibilidad de la predicción de la población de Imputados en un horizonte de 3 años

El gráfico de intervalos de confianza para las predicción de Imputados nos muestra que el intervalo de 100% de confianza es más inestable comparado con el de Condenados. Y que para el intervalo de 95% de confianza muestra una tendencia divergente que para el mes 36 alcanza un tamaño poco mayor a un 7% de la magnitud la población. Por lo que se ve que esta predicción también es de intervalo acotado y estable.

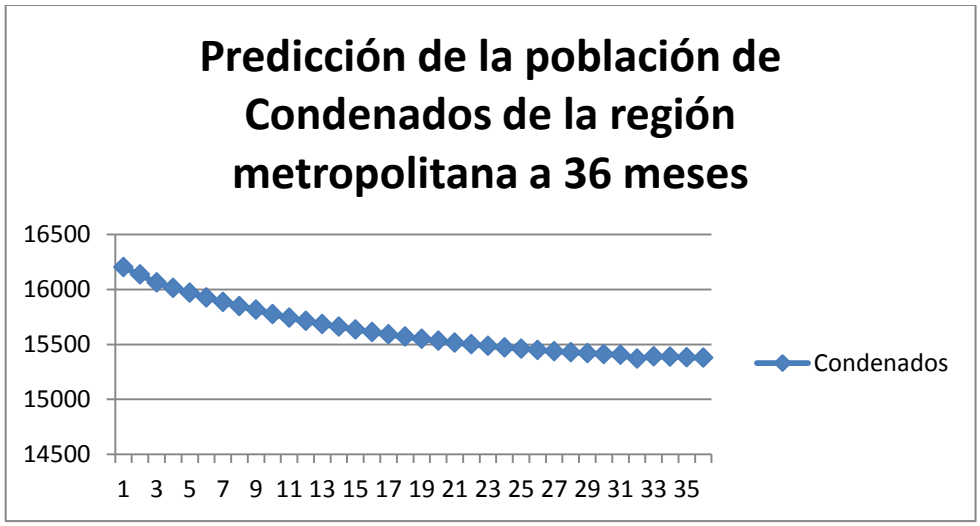

Ilustración 30. Gráfico de predicción a 36 meses de la población de condenados de la región metropolitana

Se observa que a lo largo de 3 años la población de condenados disminuye, pero que pareciera llegar a una estabilización, o, que a futuro pudiera tomar una curva ascendente.

Ahora si analizamos el gráfico de sensibilidad se observa lo siguiente:

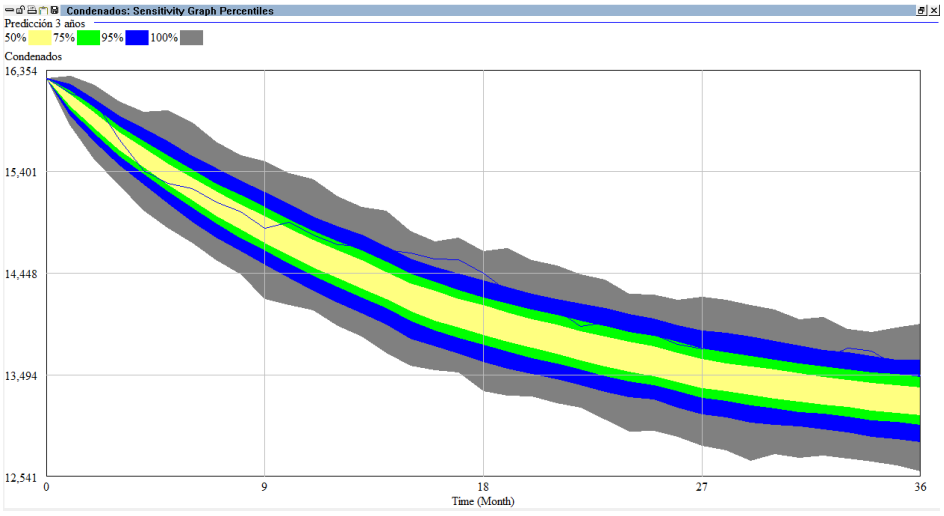

Ilustración 31. Gráfico de análisis de sensibilidad de la predicción de la población de Condenados en un horizonte de 3 años

El gráfico de intervalos de confianza para la predicción de Condenados nos muestra que en el intervalo de 100% de confianza no muestra una tendencia divergente a partir del mes 5. Y que para los 36 meses, el intervalo de 95% de confianza se estima que alcanza un tamaño aproximado de 700 que es menor a un 5% del tamaño de esta población. Por lo que se ve que esta predicción es de intervalo acotado y estable.

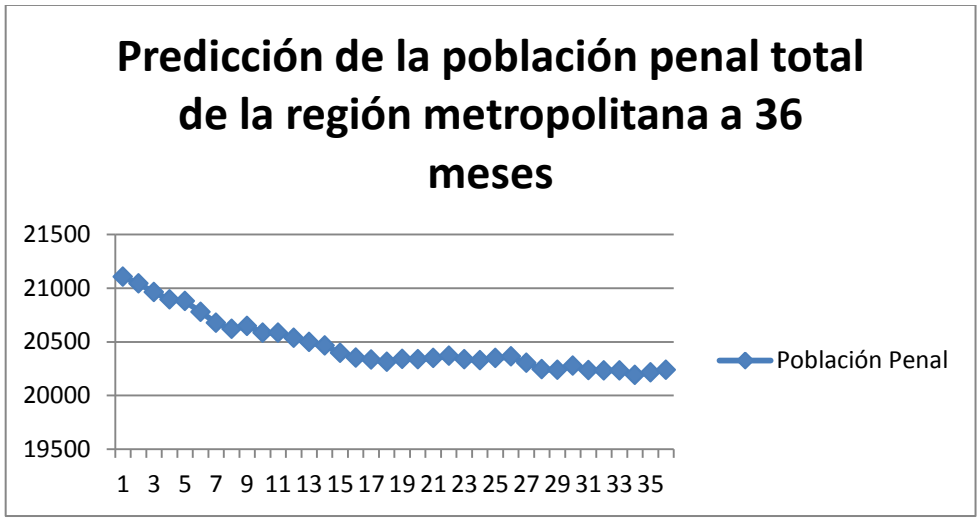

Ilustración 32. Gráfico de predicción a 36 meses de la población penal total de la región metropolitana

Al igual que el gráfico anterior, se observa una disminución a lo largo del tiempo, pero que al parecer se estabilizaría la curva de la demanda para los últimos meses.

Cabe recordar que como los resultados de cada dato predicho depende fuertemente del dato anterior, al ampliar el horizonte de tiempo de la predicción también lo hará el error de esta misma por realizarse predicciones en función de predicciones anteriores (en este caso son 36 predicciones consecutivas).

Capítulo 9

9 Evaluación del impacto en la proyección de la población penal provocado por medidas de reducción del hacinamiento en las cárceles de la región metropolitana

En este capítulo se desarrollarán dos medidas para combatir el hacinamiento en las cárceles chilenas:

- Reducir la reincidencia en un 5%.
- Deportar condenados de nacionalidad extranjera a sus países de origen.

Estas medidas fueron evaluadas en el modelo de predicción desarrollado en esta tesis, y a continuación se mostrarán los resultados.

9.1. Estrategia 1: Reducir la reincidencia en un 5%

Una posible forma de lograr esto sería el caso de que se flexibilizara la ley de forma que se dieran más arrestos domiciliarios (por ejemplo usando un brazalete electrónico) y se redujera la reincidencia.

Suponiendo que se lograra reducir la reincidencia en un 5%, se muestra la evaluación del impacto de la estrategia 1 para la proyección de las poblaciones de Imputados y Condenados.

9.1.1. Efectos de la estrategia 1 para la proyección de la población de Imputados

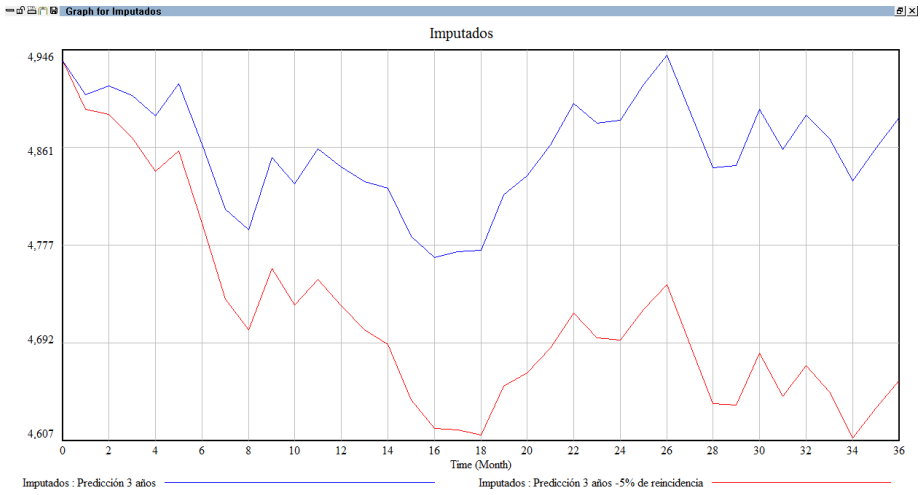

Ilustración 33. Gráfico comparativo de las proyecciones de la población de Imputados con respecto a la estrategia 1 en un horizonte de 3 años

Se observa que una variación en la reincidencia causaría un efecto más significativo en el número de imputados en el mediano plazo, ya que recién a los 12 meses difieren las predicciones en un 2,5% (120 internos aproximadamente). Esto puede deberse a que la diferencia de ingreso es pequeña en comparación a la magnitud de la población de Imputados.

Ahora con respecto a la diferencia entre las dos curvas de predicción, vemos que esta sigue aumentando. Sin embargo esta diferencia tampoco aumenta en forma lineal. Si comparamos los resultados predichos vemos que a los 12 meses teníamos una disminución de 120 imputados, mientras que para los 36 meses se habrá disminuido en aproximadamente 228 imputados, es decir, habrán diferido las predicciones en un 4,7% (que corresponde a un 89,2% de la diferencia en los 24 meses posteriores con respecto a lo que se disminuyó en los primeros 12 meses).

Esto indica que si bien el reducir la probabilidad de reincidir sigue disminuyendo la cantidad de imputados dentro de las cárceles de la región metropolitana, este efecto también comienza a estabilizarse a lo largo del tiempo, pero que sigue beneficiando en un horizonte de 3 años.

Ahora si analizamos el gráfico de sensibilidad de esta proyección vemos lo siguiente:

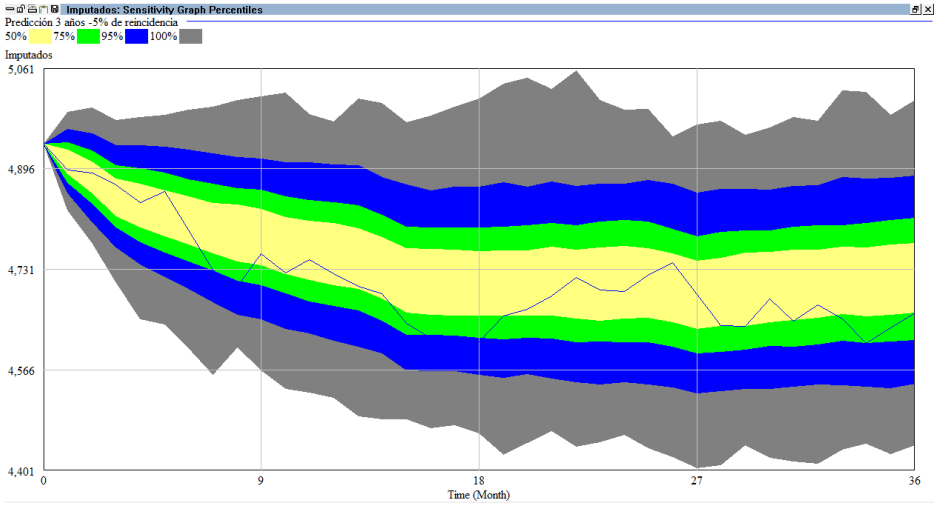

Ilustración 34. Gráfico de sensibilidad de la proyección de la población de Imputados con respecto a la estrategia 1 en un horizonte de 3 años

Se observa que se refleja el comportamiento mostrado en el gráfico anterior. Además se ve que el intervalo de confianza se mantiene acotado, lo cual indica que en el mediano plazo esta herramienta entrega resultados acotados.

9.1.2. Efectos de la estrategia 1 para la proyección de la población de Condenados

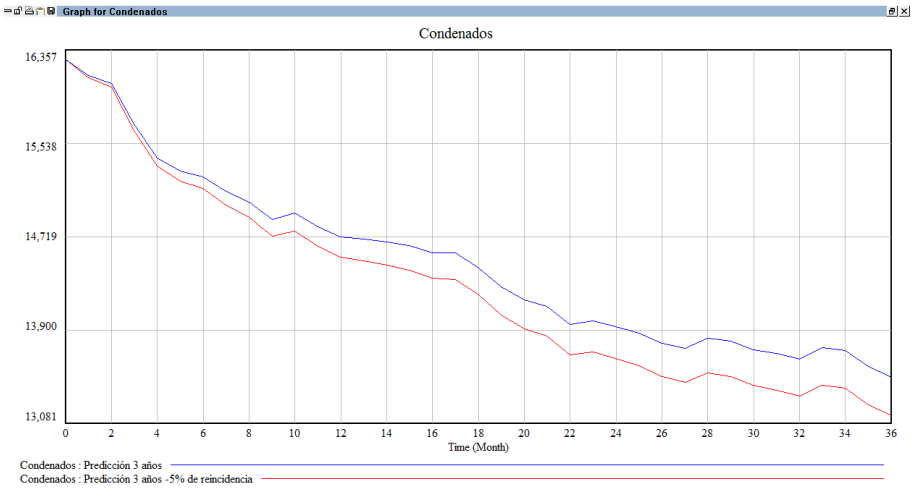

Ilustración 35. Gráfico comparativo de las proyecciones de la población de Condenados con respecto a la estrategia 1 en un horizonte de 3 años

Se observa que una variación en la reincidencia causaría un efecto más significativo en el número de condenados en el mediano plazo, ya que recién a los 12 meses difieren las predicciones en un 1,2% (179 internos aproximadamente). Esto puede deberse a que la diferencia de ingreso es pequeña en comparación a la magnitud de la población de condenados.

Ahora con respecto a la diferencia entre las dos curvas de predicción, vemos que esta sigue aumentando. Sin embargo esta diferencia tampoco aumenta en forma lineal. Si comparamos los resultados predichos vemos que a los 12 meses teníamos una disminución de 179 condenados, mientras que para los 36 meses se habrá disminuido en aproximadamente 337 condenados, es decir, habrán diferido las predicciones en un 2,5% (que corresponde a un 89,2% de la diferencia en los 24 meses posteriores con respecto a lo que se disminuyó en los primeros 12 meses).

Esto indica que si bien el reducir la probabilidad de reincidir sigue disminuyendo la cantidad de condenados dentro de las cárceles de la región metropolitana, este efecto también comienza a estabilizarse a lo largo del tiempo, pero que sigue beneficiando en un horizonte de 3 años.

Ahora si analizamos el gráfico de sensibilidad de esta proyección vemos lo siguiente:

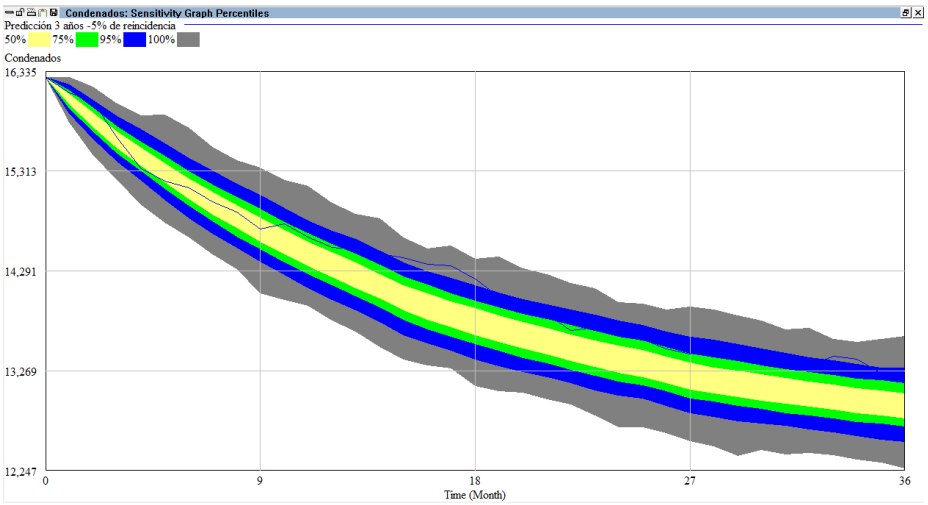

Ilustración 36. Gráfico de sensibilidad de la proyección de la población de Condenados con respecto a la estrategia 1 en un horizonte de 3 años

Se observa que se refleja el comportamiento mostrado en el gráfico anterior. Además se ve que el intervalo de confianza se mantiene acotado, lo cual indica que en el mediano plazo esta herramienta entrega resultados para la predicción acotados.

9.2. Estrategia 2: Deportar condenados de nacionalidad extranjera a sus países de origen

En agosto del año 2012 se inició un operativo para deportar a condenados de nacionalidad extranjera a sus países de procedencia [1], donde se trasladaron a 122 condenados de nacionalidad Peruana, de un total de 720 internos provenientes de 28 países.

Para medir el impacto de esta estrategia se tomó el supuesto de que en Diciembre del año 2011 (mes 0) se hubieran deportado 500 Condenados de nacionalidad extranjera y comparar los resultados con la proyección sin realizar esta acción.

A continuación se muestra la evaluación del impacto de la estrategia 2 para la proyección de las poblaciones de Imputados y Condenados.

9.2.1. Efectos de la estrategia 2 para la proyección de la población de Imputados

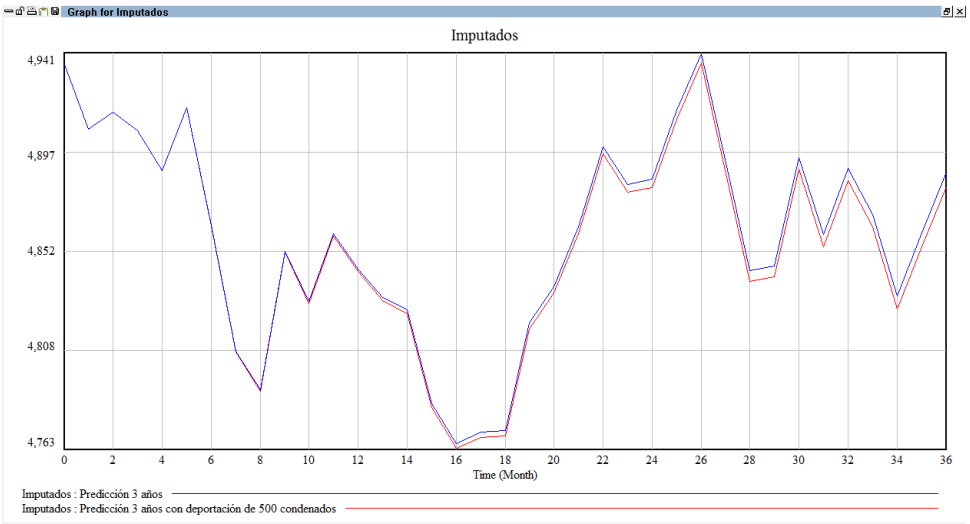

Ilustración 37. Gráfico comparativo de las proyecciones de la población de Imputados con respecto a la estrategia 2 en un horizonte de 3 años

Se observa que una variación en la reincidencia causaría un efecto casi nulo en el número de imputados en el mediano plazo, ya que recién a los 11 meses difieren las predicciones en 1 imputado, y para los 36 meses difieren en 7 imputados.

Esto indica que para la población de imputados no resulta una buena estrategia, su impacto negativo es casi nulo. Por lo que se vería opacado con respecto a la reducción de condenados que fueron deportados.

Ahora si analizamos el gráfico de sensibilidad de esta proyección vemos lo siguiente:

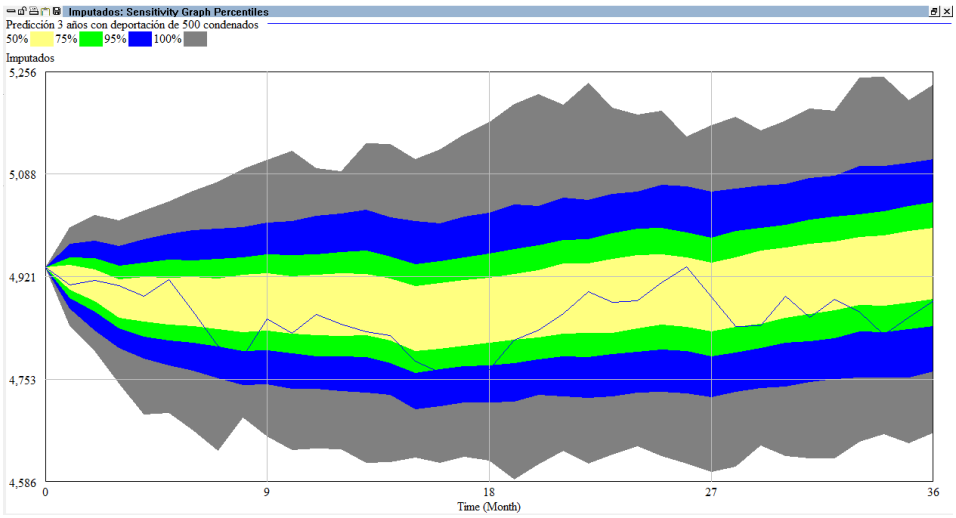

Ilustración 38. Gráfico de sensibilidad de la proyección de la población de Imputados con respecto a la estrategia 2 en un horizonte de 3 años

Se observa que se refleja el comportamiento mostrado en el gráfico anterior. Además se ve que el intervalo de confianza se mantiene acotado, lo cual indica que en el mediano plazo esta herramienta entrega resultados para la predicción acotados.

9.2.2. Efectos de la estrategia 1 para la proyección de la población de Condenados

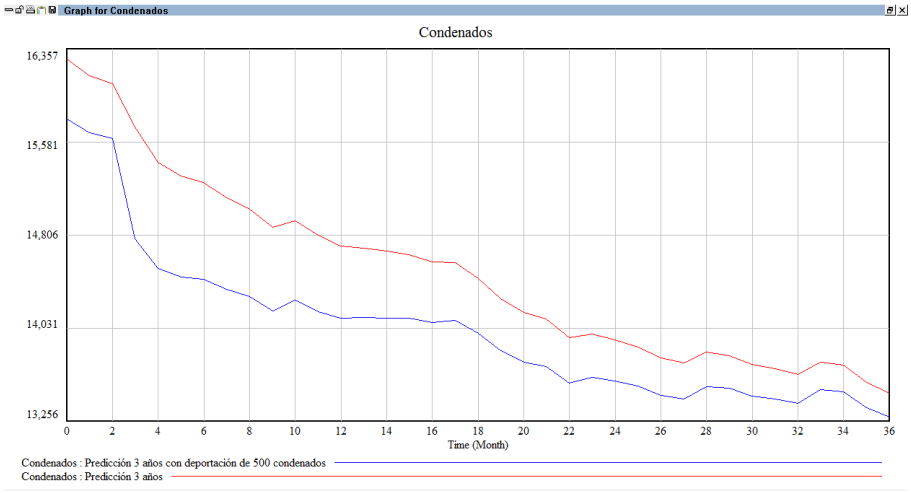

Ilustración 39. Gráfico comparativo de las proyecciones de la población de Condenados con respecto a la estrategia 2 en un horizonte de 3 años

Se observa que una variación en la reincidencia causaría un efecto más significativo en el número de condenados en el corto plazo. A 3 meses de haberse realizado la deportación, las curvas de predicción de la población de Condenados difieren en un 4,1% (604 internos aproximadamente).

Sin embargo esta diferencia disminuye a medida que avanza el tiempo. Si comparamos los resultados predichos vemos que a los 3 meses teníamos una disminución de 604 condenados, mientras que para los 36 meses diferirán en aproximadamente 198 condenados, es decir, habrán diferido las predicciones en un 1,5%.

Esto indica que si bien el deportar condenados extranjeros disminuye la cantidad de condenados dentro de las cárceles de la región metropolitana en el corto plazo, este efecto también comienza a estabilizarse a lo largo del tiempo, pero que sigue beneficiando en un horizonte de 3 años.

Ahora si analizamos el gráfico de sensibilidad de esta proyección vemos lo siguiente:

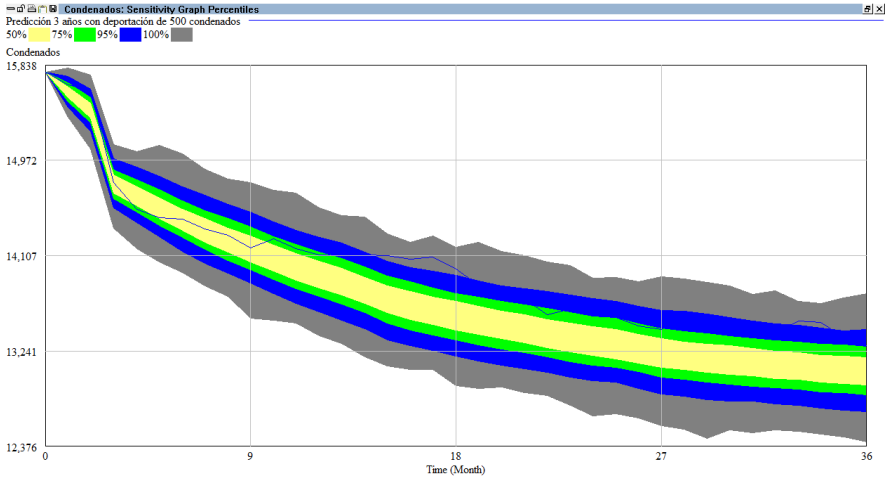

Ilustración 40. Gráfico de sensibilidad de la proyección de la población de Condenados con respecto a la estrategia 2 en un horizonte de 3 años

Se observa que se refleja el comportamiento mostrado en el gráfico anterior. Además se ve que el intervalo de confianza se mantiene acotado, lo cual indica que en el mediano plazo esta herramienta entrega resultados para la predicción acotados.

Capítulo 10

10 Conclusiones y futuros desafíos

En este capítulo se englobarán los principales resultados del desarrollo de esta tesis, como también futuras líneas de investigación a partir del trabajo realizado.

10.1. Conclusiones

- **Considerar la reincidencia dentro de este tipo de modelos entrega mejores resultados que modelar a todos los criminales de igual forma.**

  Los resultados muestran que para todos los modelos de simulación utilizados y para todos los tipos de población, el escenario con reincidencia mostró una mejor predicción que el escenario sin considerar la reincidencia. Esto se debe a que ocurre un efecto cíclico que influye en la cantidad de ingresos que los modelos sin reincidencia no lo capturan, y además el ingreso de reincidentes a los centros penitenciarios cerrados de la región metropolitana es mayor a la cantidad de primerizos que quedan en calidad de imputado o condenado a la primera.
  Sin embargo el aporte que produjo el considerar la reincidencia no fue tan significativo, por lo que también se concluye que no conviene seguir complejizando el modelo desarrollado con respecto a segmentar a la población penal por rangos etarios.

- **La simulación de dinámica de sistemas para este problema en particular funciona mejor que la simulación de eventos discretos.**

  Al comparar los resultados obtenidos de los distintos tipos de simulación se aprecia que los resultados obtenidos con la simulación de eventos discretos para predecir la demanda de población penal (Condenados e Imputados) fueron los peores.

  Además los gráficos de sensibilidad de las proyecciones de demanda de imputados y condenados obtenidas por simulación de dinámica de sistemas muestran una

estabilidad a lo largo del tiempo. Por lo que este tipo de simulación, además de entregar las predicciones con menor MAPE y MSE, entrega predicciones acotadas.

Por último el problema a evaluar corresponde a un nivel estratégico y no uno operacional ya que se está prediciendo la población penal de toda la región y no cárcel por cárcel. Esto coincide con el tipo de enfoque al cual está orientada simulación de dinámica de sistemas.

Por lo tanto se concluye que para este modelo en particular la predicción es mejor usando una simulación de dinámica de sistemas en vez de una simulación de eventos discretos.

- **Un mix de minería de datos para predicción de series de tiempo en conjunto con la simulación es una herramienta que genera resultados de confianza para predecir la demanda de la población penal de la región metropolitana.**

  Los resultados del modelo de simulación de dinámica de sistemas usando los valores medios de las variables mostraban un error de predicción MAPE para la población de condenados es menor a 3,8%, y un MAPE menor a 1,5% para la población de imputados (en un horizonte de un año). Esto es un resultado muy bueno debido a que la población de condenados de la región metropolitana en promedio un 76,56% del tamaño de la población penal total de la región metropolitana durante el año 2011, por lo que el error de predicción de los condenados sea pequeño resulta muy importante para la toma de futuras decisiones porque es la población más grande de los centros penitenciarios.

  Con esto se puede deducir que si se contara con una buena predicción de un mediano plazo de la variable Aprehendidos mayores de edad en la región metropolitana se pueden obtener resultados de confianza para tomar decisiones de asignación de recursos futuros a esa población penal. Como por ejemplo asignar eficientemente la cantidad de nuevos gendarmes necesarios para cuidar a este tipo de internos en la región metropolitana, construir una nueva cárcel en la región anticipándose a la demanda de internos, asignar recursos eficientemente, etc. Y dado que la predicción de los detenidos mayores de edad en cada mes se realiza con el valor real hace 24 meses atrás más las variables dummies correspondientes a Enero, Febrero y Septiembre, se contarán con datos reales para poder predecir un horizonte de 2 años de esta serie de tiempo, por lo que se piensa que con este modelo se podría tomar estas decisiones. De hecho, un año es un horizonte de tiempo en el cual se puede

construir una cárcel (por ejemplo la cárcel de Parral comenzará construirse en Septiembre del año 2013 y se entregará la obra en Julio del año 2014 [30]. por lo que este modelo de predicción permitiría planificar mejor la capacidad de internos y el espacio designado para Condenados y para Imputados (ya que tienen que estar separados dentro del recinto penal).

Además se observó que los intervalos de confianza de los gráficos de sensibilidad de la predicción de la población penal se mantenían acotados tanto para la población de condenados como para la de imputados. Por lo que sumándole el hecho de que el error de predicción MAPE en el horizonte de un año para la población penal total fue menor a un 3%, se puede concluir que se obtienen resultados consistentes para el corto plazo.

Sin embargo desde un punto de vista económico se debería replantear la predicción de los Aprehendidos mayores de edad para un horizonte de un mediano plazo porque para un horizonte mayor de predicción se deben considerar variables explicativas como por ejemplo variables sociodemográficas, tasa de desempleo, etc…

- **Reducir la reincidencia y deportar condenados extranjeros son estrategias con soluciones temporales**

  El reducir la reincidencia es resulta ser efectiva, teniendo mayor impacto durante el primer año, pero que para el final del tercer año el efecto va disminuyendo y se estabiliza, por lo que esta estrategia tiene un horizonte de efectividad aproximado de 3 años.

  Por otra parte, el deportar condenados extranjeros es una solución instantánea que permite reducir la población en el corto plazo. Pero que en el mediano plazo también se estabiliza, y resulta ser una peor estrategia que reducir la reincidencia.

  Por lo tanto se deduce que una estrategia más sostenible en el tiempo sería disminuir constantemente la probabilidad de reincidencia.

- **El modelo utilizado permite realizar análisis de situaciones “what-if”**

  Este punto quedó demostrado al analizar las estrategias descritas anteriormente. Pero estas estrategias se definieron debido al alcance de las acciones a realizar de parte de Gendarmería de Chile.

Ahora si se involucraran a otras instituciones participes del proceso penal como Carabinero de Chile, Policía de Investigaciones o al Ministerio Público se podrían evaluar diferentes estrategias mixtas (mejorar la efectividad de Aprehensiones de Carabineros de Chile y/o de Policía de Investigaciones, o mejorar la efectividad de condena de parte de la fiscalía por ejemplo). Y al modelar el problema en base a un modelo de flujo, este permite la flexibilidad de remodelarlo en función de otros procesos que impactan como un flujo hacia el sistema penitenciario.

10.2. Futuros desafíos

- **Mejorar la predicción de la serie de tiempo utilizada como input en el modelo**

  Como se explicó en las conclusiones, el predecir una serie de tiempo de los aprehendidos mayores de edad de la región metropolitana sin variables explicativas (es decir, solo con variables históricas) para el mediano plazo, según el punto de vista económico, pierde precisión en los resultados. Por lo que si se quiere mayor consistencia en los resultados de predicción a un mediano plazo se debe profundizar más en esta etapa del modelo.

  Además, si se segmenta a la población como se analizó en el punto anterior, puede que sea necesario segmentar los ingresos de los detenidos sin antecedentes penales, por lo que podrían necesitarse más series de tiempo en vez de una sola.

- **Predecir el comportamiento futuro de los parámetros a través del tiempo**

  En esta tesis se mantuvieron los parámetros constantes durante las predicciones en el mediano plazo, pero en la realidad se sabe que eso difícilmente ocurre. Por lo que habría que estudiar el comportamiento de los parámetros y/o variables para obtener predicciones más confiables.

- **Evaluar estrategias mixtas con distintas instituciones involucradas en el proceso penal**

  Como se dijo en las conclusiones, el flexibilizar el modelo permitiría evaluar nuevas estrategias con el fin de disminuir el nivel de sobrepoblación penal en las cárceles. Y con más entes participantes se amplía el mix de estrategias a evaluar y así poder determinar una estrategia más óptima y sin limitaciones para ejercerla.

- **Flexibilizar el modelo para predecir la población penal del sistema abierto**

  Dado que se tienen que destinar recursos para atender a la población penal del sistema abierto (profesionales, oficinas de atención, visitas a terreno, infraestructura de apoyo, etc.). Además de que la población penal del sistema abierto supera a la población penal del sistema cerrado (50 mil vs 47 mil aproximadamente) [17] por lo que su inclusión podría ser necesaria para los fines de Gendarmería de Chile

Bibliografía

[1] AMBITO.COM. Chile inició deportación de presos extranjeros con el traslado de 122 peruanos. [En línea] http://www.ambito.com/noticia.asp?id=648887. [Último acceso: 18 NOVIEMBRE 2013].

[2] Baeza, A. Sobrepoblación penitenciaria cae de 54% a 22% en tres años y llega a nivel más bajo desde 2010. La Tercera en Internet. 25 abril 2013. [En línea] http://www.latercera.com/noticia/nacional/2013/04/680-520132-9-sobrepoblacion-penitenciaria-cae-de-54-a-22-en-tres-anos-y-llega-a-nivel-mas.shtml. [Último acceso: 17 octubre 2013].

[3] Baker, J. y Lattimore, P. Forecasting Demand Using Survival Modeling: An Appication To US Prisons. Australasian Journal of Information Systems; Vol 2, No 1 (1994).

[4] Barnett, A. Prison Populations: A Projection Model. Operations Research, Vol. 35, No. 1 (Jan. - Feb., 1987), pp. 18-34.

[5] Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. Volume 52, Issue 3. pp 345-370.

[6] Bravo, C. Minería de Datos aplicada a Teoría de Juegos. Teoría y Aplicación a la Industria Financiera. Tesis (Magíster en Gestión de Operaciones, Ingeniero Civil Industrial). Santiago, Chile. Universidad de Chile. 2008.

[7] CARABINEROS DE CHILE. *2013.* [en línea] http://www.carabineros.cl [consulta: 22 junio 2014].

[8] Cristiannini, N. y Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-Based Methods. Inglaterra. Cambridge University Press. 2003. 190p.

[9] Chatfield, C. Model Uncertainity, Data Mining and Statistical Inference. Journal of the Royal Statistical Society. 158:419-466. 1995.

[10] CHILE. *Gendarmería de Chile. 1979. Ley 2.859: Fija Ley Orgánica de Gendarmería de Chile, 12 septiembre 1979.*

[11] Defensoría de los Habitantes de la República de Costa Rica. 2012. COSTA RICA: Defensoría de los Habitantes, Defensa Pública del Poder Judicial y Mecanismo Nacional de Prevención de la Tortura piden acciones urgentes al Gobierno. [en línea] <http://www.portalfio.org/inicio/noticias/item/10950-costa-rica-defensor%C3%ADa-de-los-habitantes-defensa-p%C3%BAblica-del-poder-judicial-y-mecanismo-nacional-de-prevenci%C3%B3n-de-la-tortura-piden-acciones-urgentes-al-gobierno.html> [consulta: 22 junio 2014].

[12] EL MERCURIO. 2011. Incendio en Cárcel de San Miguel deja 81 reos fallecidos y obliga a evacuar a otros 200. [En línea] http://www.emol.com/noticias/nacional/2010/12/08/451604/incendio-en-carcel-de-san-miguel-deja-81-reos-fallecidos-y-obliga-a-evacuar-a-otros-200.html. [consulta: 17 octubre 2013].

[13] Fayyad, U., Piatetsky-Schapiro, G. y Smyth, P. From Data Mining to Knowledge Discovery in Databases. Communications of the ACM. 39(11):24-26. 1996.

[14] Foix, C. 2006. Proyección del precio del cobre: ¿Herramientas de inteligencia computacional o series de tiempo?. Tesis (Magíster en Gestión de Operaciones). Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

[15] Frey, A. Población Penal Y Mercado Laboral: Un Modelo Empírico para el Período 1982-2002 en la Región Metropolitana. Estudio de Caso nº76 (Magíster en Gestión de Políticas Públicas). Santiago, Chile. Universidad de Chile. 2004.

[16] Gazmuri, P. [2012]. Clase nº1 Modelos de Simulación Discreta [Diapositivas] http://www.siding.puc.cl.

[17] GENDARMERIA DE CHILE. 2014. [en línea] http://www.gendarmeria.gob.cl [consulta: 22 junio 2014].

[18] Hansen, J., Macdonald, J. y Nelson, R. Some Evidence on Forecasting Time-Series with Support Vector Machines. The Journal of the Operational Research Society, Vol. 57, No 9 (Sep., 2006), pp. 1053-1063.

[19] Hastie, T., Tibshirani, R. y Friedman, J. The Elements of Statistical Learning - Data Mining, Inference and Prediction. EE. UU. Springer, 3ª edición, 2003. 552p.

[20] Jacobs, D. y Helms, R. Toward a Political Model of Incarceration: A Time-Series Examination of Multiple Explanations for Prison Admission Rates. American Journal of Sociology Vol. 102, No. 2(Sep., 1996), pp. 323-357.

[21] Jiménez, D. 2011. Análisis y Pronósticos de Demanda para Telefonía Móvil. Tesis (Ingeniera Civil Industrial, Magíster en Gestión de Operaciones). Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. 63p.

[22] Jory, M. 2007. Predicción de las variaciones de costos para proyectos de construcción utilizando redes neuronales. Memoria (Ingeniero Civil Industrial). Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. pp. 40-64.

[23] Labrín, S. Hacinamiento en cárceles: Fisco ha pagado más de $5 mil millones a concesionarias. La Tercera en Internet. 23 julio 2012. [En línea] http://diario.latercera.com/2012/07/23/01/contenido/pais/31-114395-9-hacinamiento-en-carceles-fisco-ha-pagado-mas-de--5-mil-millones-a-concesionarias.shtml. [consulta: 17 octubre 2013].

[24] Lochner, L. Education, Work and Crime: A Human Capital Approach. International Economic Review, Vol. 45, No. 3 (Aug., 2004), pp. 811-843.

[25] Mackinson, M. y Glick, N. Data Mining and Knowledge Discovery in Databases - An Overview. Australian & New Zealand Journal of Statistics. 41(3):255-275. 1999.

[26] Macleod, J. Modelling crime and offending: recent developments in England and Wales: Section C. London: Home Office (Ocassional paper No.80 Section C).

[27] Marianov, V. Localización y Dimensionamiento de Establecimientos Penitenciarios. Santiago, Chile, Pontificia Universidad Católica de Chile, DICTUC. 2001. pp 37-42.

[28] Matich, D. 2001. Redes Neuronales: Conceptos Básicos y Aplicaciones. [en línea] <http://www.itescam.edu.mx/principal/sylabus/fpdb/recursos/r101038.PDF> [consulta: 22 junio 2014].

[29] McDowall, R. Prison Overcrowding: Finding Successful Policies to Manage Capacity Utilization. Tesis (Master of Philosophy in System Dynamics). Bergen, Noruega. University of Bergen. 2010.

[30] MOLINACHILE. 2013. En Septiembre comenzará construcción de la cárcel de Parral. [En línea] http://www.molinachile.cl/en-septiembre-comenzara-construccion-de-la-carcel-de-parral [consulta: 19 noviembre 2013].

[31] Myatt, G. Making Sense of Data. New Jersey, EE. UU. John Wiley & Sons Inc. 2007. 293p.

[32] New Mexico Sentencing Comission Staff, 2010. New Mexico Prison Population Forecast: FY 2011-FY 2020. [en línea] http://nmsc.unm.edu/reports/2010/new-mexico-prison-population-forecast-fy-2011-fy-2020.pdf [consulta: 6 Noviembre 2013].

[33] Oficina de las Naciones Unidas Contra la Droga y el Delito. 2010. El sistema penitenciario: Medidas privativas y no privativas de libertad. [en línea] <http://www.unodc.org/documents/justice-and-prison-reform/crimeprevention/The_Prison_System_Spanish.pdf> [consulta: 22 junio 2014].

[34] POLICÍA DE INVESTIGACIONES. *2014.* [en línea] http://www.policia.cl [consulta: 22 junio 2014].

[35] Rumelhart, D. y McClelland, J. y el grupo de investigación PDP. Parallel Distributed Processing: Explorations on the Microstructure of Cognition. MIT Press. Cambridge, EE. UU. 1986.560p.

[36] Schölkopf, B. y Smola, A. J. A Tutorial on Support Vector Regression. Statistics and Computing. Volume 14, Issue 3, pp199-222.

[37] Schölkopf, B. y Smola, A. J. Learning with Kernels .Support Vector Machines, Regularization, Optimization, and Beyond, 644 pages, MIT Press, Cambridge, MA, USA, (2002).

[38] Seber, G. y Lee, A. Linear Regression: Estimation and Distribution Theory. En: Linear Regression Analysis. 2ª Edición ed. New Jersey. John Wiley & Sons Inc. 2003. pp 35-96.

[39] Shannon, R. y Johannes, J. D. Systems simulation: the art and science. IEEE Transactions on Systems, Man and Cybernetics 6(10).1976. pp. 723-724.

[40] Shannon, R. y Johannes, J. D. From Systm Dynamics and Discrete Events to Practical Agent Based Modeling: Reasons, Techniques, Tools.

[41] TELESUR TV. *2011. Hacinamiento, mayo problema en cárceles chilenas. [en línea] http://http://multimedia.telesurtv.net/media/telesur.video.web/telesur-web/#!es/video/hacinamiento-mayor-problema-en-carceles-chilenas/ [consulta: 17 octubre 2013].*

[42] Vapnik, V. y Cortes, C. Support Vector Networks. Machine Learning. 20:1-25. 1995.

[43] Vapnik, V., Smola, A., Kaufman, L., Burges, C. y Drucker, H. 1997. Support Vector Regression Machines. En: Mozer, M., Jordan, M. y Petsche, T. (Eds.). Advances in Neural Information Processing Systems 9. Estados Unidos. Massachusetts Institute of Technology. pp. 155-161.

[44] Werbos, P. The Roots of Backpropagation. Wiley-IEEE. Nueva York, EE. UU. 1994. 319p.

[45] Zhang, P. Avoiding Pitfalls in Neural Network Research. IEEE Transactions on Systems, Man and Cybernetics: Part C - Applications & Reviews. 37(1):3-16. 2007.

ANEXOS

Anexo A: Información obtenida

La información para el desarrollo de esta tesis se obtuvo a través de 3 fuentes: Información obtenida por convenio de traspaso de información entre el Centro de Análisis y Modelamiento en Seguridad (CEAMOS) y Gendarmería de Chile, solicitud de información a Carabineros de Chile mediante ley de transparencia, y a través de preguntas a funcionarios de Carabineros de Chile apelando a su juicio de expertos.

A continuación se detallará la información obtenida en cada fuente:

- Convenio de traspaso de información entre CEAMOS y Gendarmería de Chile:
    - Cantidad de internos en calidad de Condenados dentro de una cárcel de la región metropolitana en el último día del mes desde Enero del año 2005 hasta Marzo del año 2012.
    - Cantidad de internos en calidad de Imputados dentro de una cárcel de la región metropolitana en el último día del mes desde Enero del año 2005 hasta Marzo del año 2012.
    - Cantidad de personas que ingresaron a una cárcel de la región metropolitana en calidad de Condenado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2011.
    - Cantidad de personas que ingresaron a una cárcel de la región metropolitana en calidad de Imputado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2011.
    - Cantidad de personas que ingresaron a una cárcel de la región metropolitana desde un centro del Servicio Nacional de Menores (SENAME) al cumplir mayoría de edad durante cada mes desde Enero del año 2005 hasta Diciembre del año 2011.
    - Cantidad de presos que ingresaron a una cárcel de la región metropolitana trasladados desde otro centro penitenciario ubicado en otra región en calidad de Condenado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2012.
    - Cantidad de presos que ingresaron a una cárcel de la región metropolitana trasladados desde otro centro penitenciario ubicado en otra región en calidad de Imputado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2012.
    - Cantidad de presos que ingresaron a una cárcel de la región metropolitana trasladados desde otro centro penitenciario ubicado en otra región en calidad

de Condenado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2012.
    - Cantidad de presos que egresaron de una cárcel de la región metropolitana trasladados hacia otro centro penitenciario ubicado en otra región en calidad de Condenado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2012.
    - Cantidad de presos que egresaron de una cárcel de la región metropolitana trasladados hacia otro centro penitenciario ubicado en otra región en calidad de Imputado durante cada mes desde Enero del año 2005 hasta Diciembre del año 2012.
    - Cantidad de presos que egresaron de una cárcel de la región metropolitana, clasificados por sus respectivos motivos de egreso (Cumplimiento de condena, libertad condicional, muertes, etc…).
    - Cantidad de presos en calidad de Imputados de una cárcel de la región metropolitana que cambian de estado a Condenados durante cada mes desde Junio del año 2009 hasta Agosto del año 2011.
- Solicitud de Información a Carabineros de Chile mediante ley de transparencia:
    - Cantidad de aprehendidos por mes dentro de la región metropolitana que son mayores de edad desde Enero del año 2005 hasta Diciembre del año 2011.
    - Cantidad de aprehendidos por mes dentro de la región metropolitana que son menores de edad desde Enero del año 2005 hasta Diciembre del año 2012.
- Preguntas a funcionarios de Carabineros de Chile apelando a su juicio de expertos:
    - Porcentaje promedio de aprehendidos mayores de edad mensuales de la región metropolitana que no poseen antecedentes penales.
    - Tamaño de la población mayor de edad de la región metropolitana que cuenta con antecedentes penales y que está en libertad en Diciembre del año 2010.
    - Porcentaje promedio mensual de aprehendidos mayores de edad de la región metropolitana que no cuentan con antecedentes penales que efectivamente ingresan a un centro penitenciario de la región metropolitana.
    - Porcentaje promedio mensual de aprehendidos mayores de edad de la región metropolitana que no cuentan con antecedentes penales que no ingresan a un centro penitenciario cerrado (por ejemplo quedan con alguna pena alternativa como brazalete electrónico) pero que quedan con antecedentes penales.

Cabe destacar que las preguntas a funcionarios se realizaron debido a que no se pudo conseguir esa información a través de la ley de transparencia.

Anexo B: Cálculo de distribución aleatoria de cada variable del modelo de predicción de la población penal

B.1 Distribución aleatoria de la variable $PITC_t$

En este caso si bien se muestran los datos estables desde Octubre 2007 hasta Diciembre 2010 se aprecia un cambio en la serie de tiempo a partir de Enero 2011 en adelante.

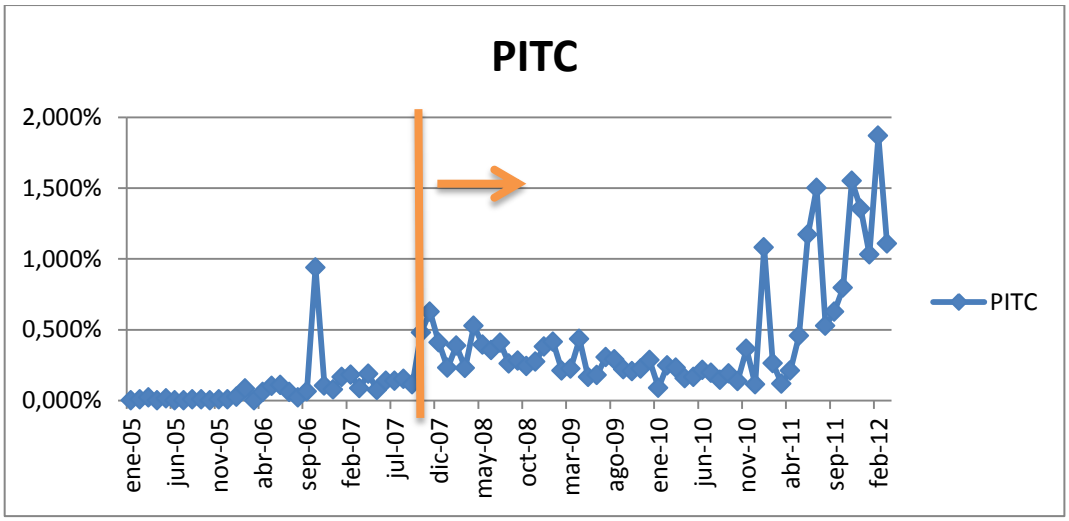

Ilustración 41. Gráfico de la variable PITC a través del tiempo

En este caso si comparamos este gráfico con el del índice ITC vemos que si bien existe esa tendencia presente en el gráfico anterior, esta se estabiliza a partir desde mayo 2012 en adelante.

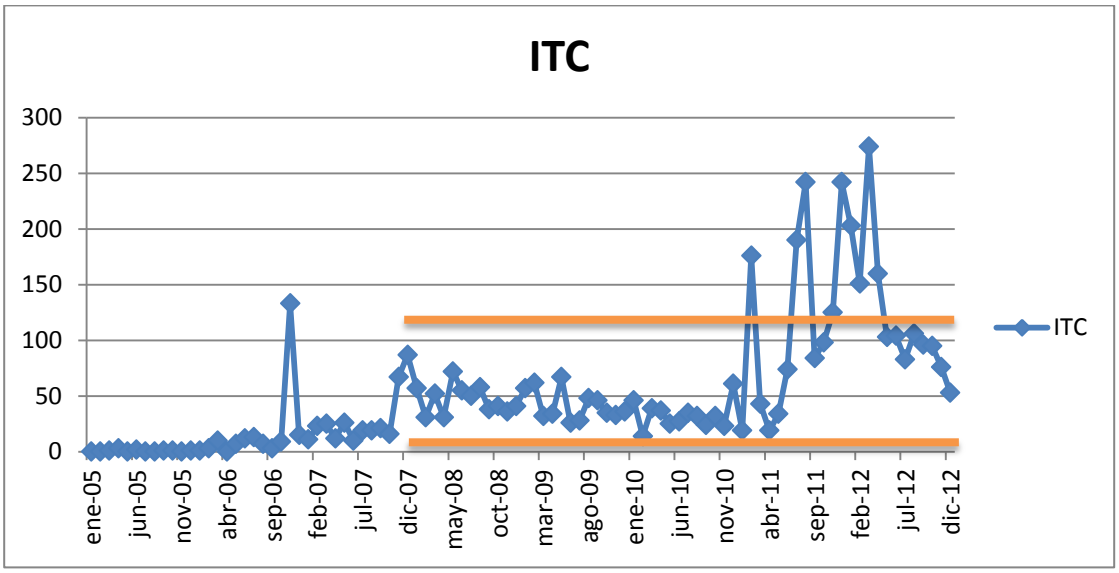

Ilustración 42. Gráfico de la variable ITC a través del tiempo

Por lo tanto se consideraron los datos del índice PITC en el rango de tiempo entre Septiembre 2007 y Diciembre 2010 tomando el supuesto de que los datos desde Enero 2011 y Marzo 2012 son outliers porque en el gráfico del índice ITC se estabilizan a partir de Mayo del 2012.

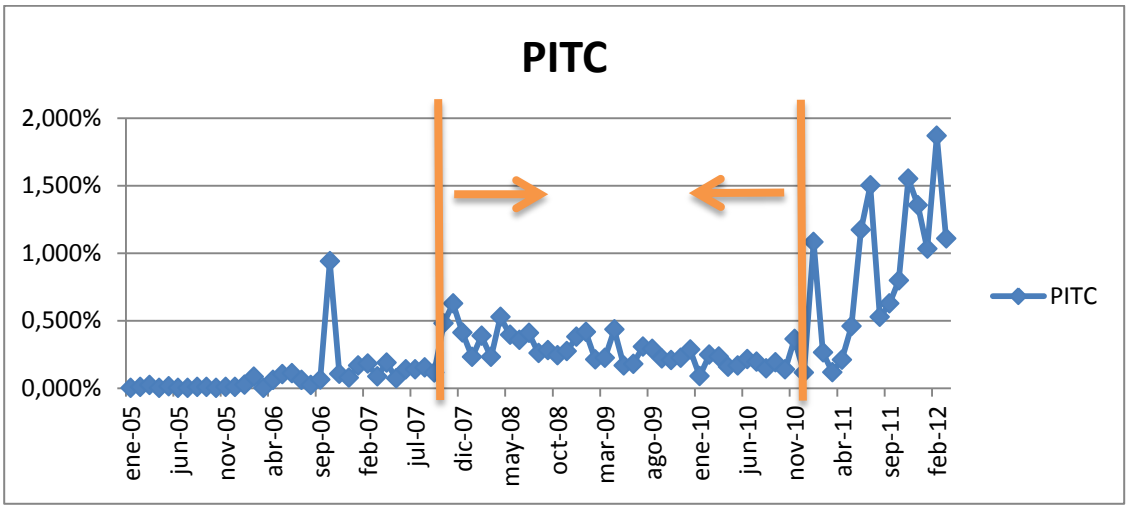

Ilustración 43. Gráfico de la variable PITC a través del tiempo donde se muestra el rango de datos a considerar para el estudio

Cantidad de datos	40
Media	0,275%
Desviación estándar	0,00121449
Total Outliers sobre 2σ	2

Tabla 22. Análisis de outliers de los datos de la variable PITC

La tabla anterior muestra que existen 2 outliers para un intervalo de ±2σ que corresponden a los datos de Noviembre del 2007 y Abril del 2008. Por lo que se obtuvo la distribución de la variable aleatoria sin considerar esos outliers mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	Weibull(0,00292; 2,83)
p-valor test Chi-Cuadrado	<0,005
p-valor test Kolmogorov-Smirnov	>0,15

Tabla 23. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PITC

Distribución	Error Cuadrático
Weibull	0,00245
Normal	0,00368
Beta	0,00412
Gamma	0,00903
Erlang	0,0113
Lognormal	0,0118
Triangular	0,119
Exponential	0,2
Uniform	0,242

Tabla 24. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PITC

Si bien el p-valor del test Kolmogorov-Smirnov es mayor a 0,15 para la distribución Weibull, el p-valor del test Chi-Cuadrado es menor a 0,005 y también Weibull es la distribución con menor error cuadrático.

B.2 Distribución aleatoria de la variable $PITI_t$

Dado el gráfico a continuación se tomaron los datos a partir de Enero del 2008 ya que a pesar de que se aprecia una variabilidad en los datos, estos tienen una menor amplitud real que en el caso anterior.

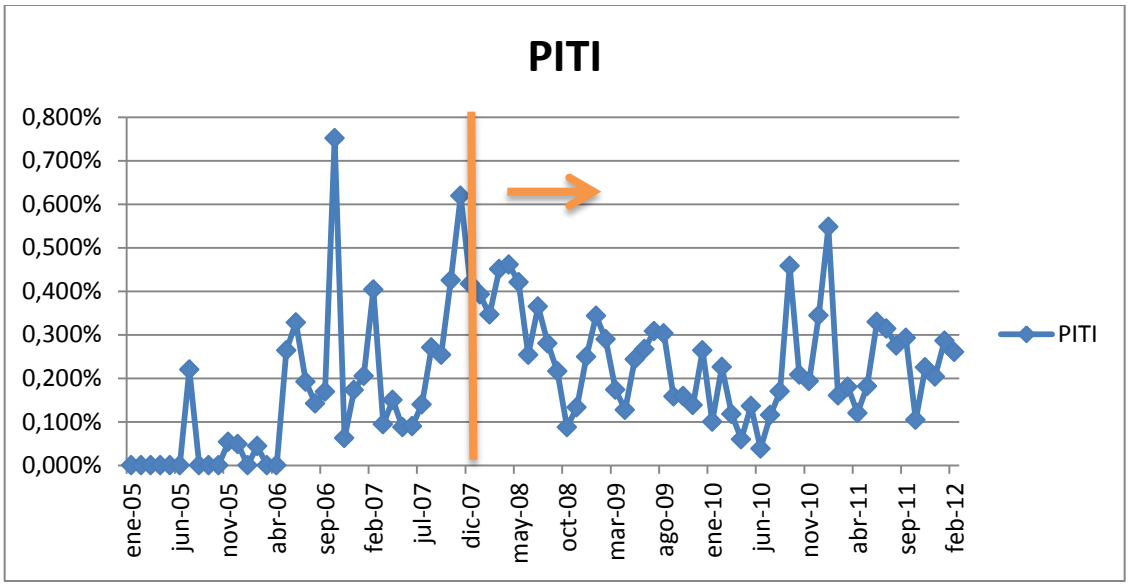

Ilustración 44. Gráfico de la variable PITI a través del tiempo donde se muestra el rango de datos a considerar para el estudio

Cantidad de datos	51
Media	0,245%
Desviación Estándar.	0,116%
Total Outliers sobre 2σ	1

Tabla 25. Análisis de outliers de los datos de la variable PITI

La tabla anterior muestra que existe 1 outlier para un intervalo de ±2σ que corresponde al dato de Febrero del 2011. Por lo que se obtuvo la distribución de la variable aleatoria sin considerar ese outlier mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	Weibull(0,0027; 2,39)
p-valor test Chi-Cuadrado	<0,005
p-valor test Kolmogorov-Smirnov	>0,15

Tabla 26. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PITI

Distribución	Error Cuadrático
Weibull	0,00656
Beta	0,00709
Erlang	0,00776
Normal	0,00925
Gamma	0,00979
Lognormal	0,0123
Triangular	0,0886
Exponential	0,0971
Uniform	0,173

Tabla 27. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PITI

Si bien el p-valor del test Kolmogorov-Smirnov es mayor a 0,15 para la distribución Weibull, el p-valor del test Chi-Cuadrado es menor a 0,005 y también Weibull es la distribución con menor error cuadrático.

B.3 Distribución aleatoria de la variable PETCt

Dado el gráfico a continuación se tomaron los datos a partir de Enero del 2008 ya que se aprecia un rango de datos más estable a excepción de un valor de septiembre del año 2010.

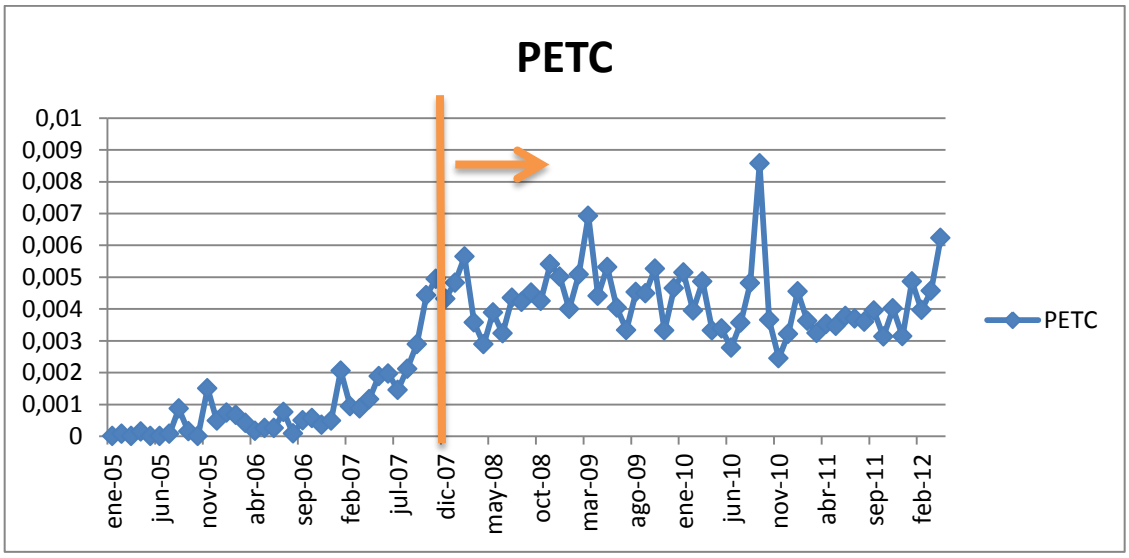

Ilustración 45. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PITI

Cantidad de datos	52
Media	0,423%
Desviación Estándar.	0,108%
Total Outliers sobre 2σ	2

Tabla 28. Análisis de outliers de los datos de la variable PETC

La tabla anterior muestra que existen 2 outliers para un intervalo de ±2σ que corresponde a los datos de Marzo del 2009 y de Septiembre del 2010. Por lo que se obtuvo la distribución de la variable aleatoria sin considerar esos dos outliers mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	Beta(14,4; 20,7873)
p-valor test Chi-Cuadrado	<0,005
p-valor test Kolmogorov-Smirnov	>0,15

Tabla 29. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PETC

Distribución	Error Cuadrático
Beta	0,000813
Erlang	0,00125
Gamma	0,0014
Normal	0,00141
Lognormal	0,00309
Weibull	0,0104
Triangular	0,163
Uniform	0,317
Exponential	0,369

Tabla 30. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PETC

Si bien el p-valor del test Kolmogorov-Smirnov es mayor a 0,15 para la distribución Beta, el p-valor del test Chi-Cuadrado es menor a 0,005 y también Beta es la distribución con menor error cuadrático.

B.4 Distribución aleatoria de la variable PETIt

Dado el gráfico a continuación se decidió tomar los datos a partir de Enero del 2008 ya que se aprecia un rango de datos más estable.

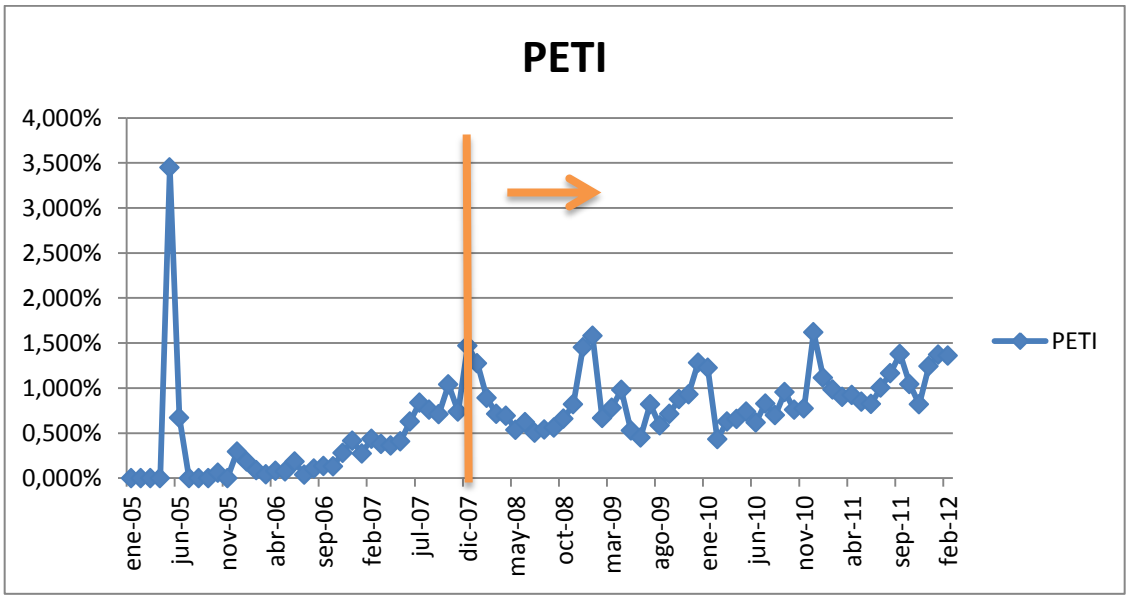

Ilustración 46. Gráfico de la variable PETI a través del tiempo donde se muestra el rango de datos a considerar para el estudio

Cantidad de datos	51
Media	0,899%
Desviación Estándar.	0,310%
Total Outliers sobre 2σ	2

Tabla 31. Análisis de outliers de los datos de la variable PETI

La tabla anterior muestra que existen 2 outliers para un intervalo de ±2σ que corresponde a los datos de Febrero del 2009 y de Enero del 2011. Por lo que se obtuvo la distribución de la variable aleatoria sin considerar esos dos outliers mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	Lognormal(0,0087; 0,00283)
p-valor test Chi-Cuadrado	0,079
p-valor test Kolmogorov-Smirnov	>0,15

Tabla 32. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PETI

Distribución	Error Cuadrático
Lognormal	0,00923
Erlang	0,0124
Gamma	0,0125
Beta	0,0199
Weibull	0,0244
Normal	0,0266
Triangular	0,0353
Uniform	0,14
Exponential	0,186

Tabla 33. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PETI

Si bien el p-valor del test Kolmogorov-Smirnov es mayor a 0,15 para la distribución Lognormal, el p-valor del test Chi-Cuadrado es 0,079 lo que si bien no es menor a 0,05 se acerca. Esto podría deberse a la poca cantidad de datos con la que se cuenta para poder obtener un ajuste con mejor exactitud.

Además la distribución Lognormal es la distribución con menor error cuadrático.

B.5 Distribución aleatoria de la variable CEt

Si bien en este caso no se cuenta con muchos datos, se puede apreciar que a partir del dato de Septiembre del año 2009 se observa un conjunto de datos más acotados que si los tomamos con anterioridad. Por lo que consideraremos los datos que se tengan de ahí en adelante.

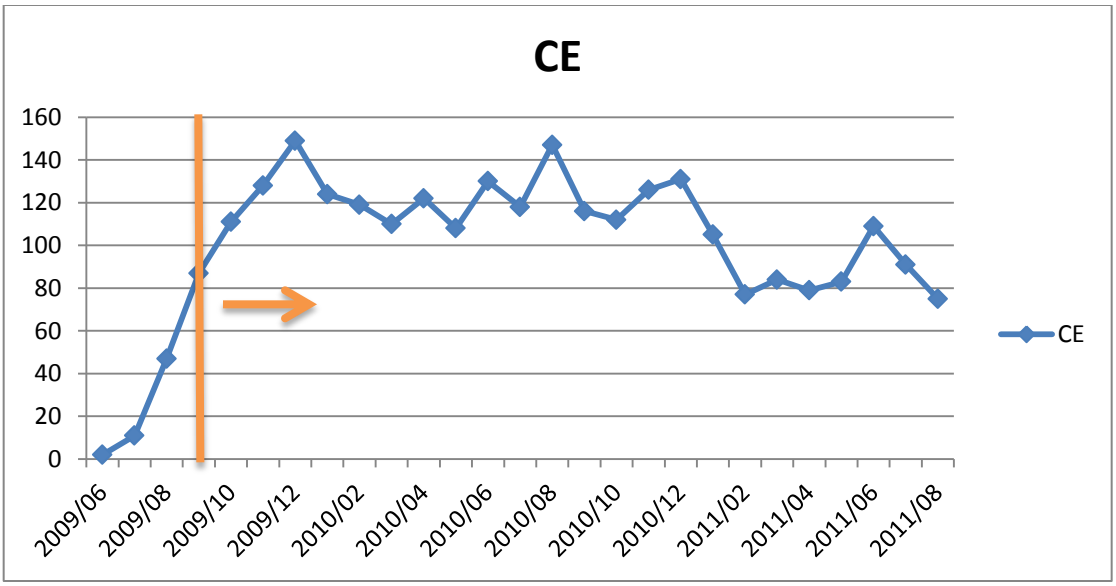

Ilustración 47. Gráfico de la variable CE a través del tiempo donde se muestra el rango de datos a considerar para el estudio

Cantidad de datos	24
Media	110,04
Desviación Estándar.	9,78
Total Outliers sobre 2σ	10

Tabla 34. Análisis de outliers de los datos de la variable CE

En esta ocasión se cuenta con pocos datos (24) y si se considera un intervalo de ±2σ nos encontramos con 10 outliers.

Ahora si se considera un intervalo de ±3σ aún existirían 5 outliers, lo cual no deja de ser más de un 20% de la cantidad de datos con los que se cuenta. Por lo tanto se considerarán todos los datos para obtener la distribución de la variable aleatoria mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	74,5 + 75 * Beta(0,857; 0,957)
p-valor test Chi-Cuadrado	0,0103

Tabla 35. Resultados del Test Chi-Cuadrado a los datos de la variable CE

Distribución	Error Cuadrático
Beta	0,0279
Uniform	0,0283
Normal	0,0289
Weibull	0,0308
Exponential	0,031
Erlang	0,031
Gamma	0,0312
Triangular	0,0317
Lognormal	0,0337
Poisson	0,0362

Tabla 36. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable CE

El p-valor del test Chi-Cuadrado es 0,0103 que es menor a 0,05 se acerca. Y como además el error cuadrático de la distribución Beta es la menor de todas la aceptamos como la mejor distribución para esta ocasión.

B.6 Distribución aleatoria de la variable PEIt

Este gráfico presenta una estabilidad visual de los datos a partir del mes de Agosto del 2008 en adelante, por lo que consideraremos los datos desde esa fecha.

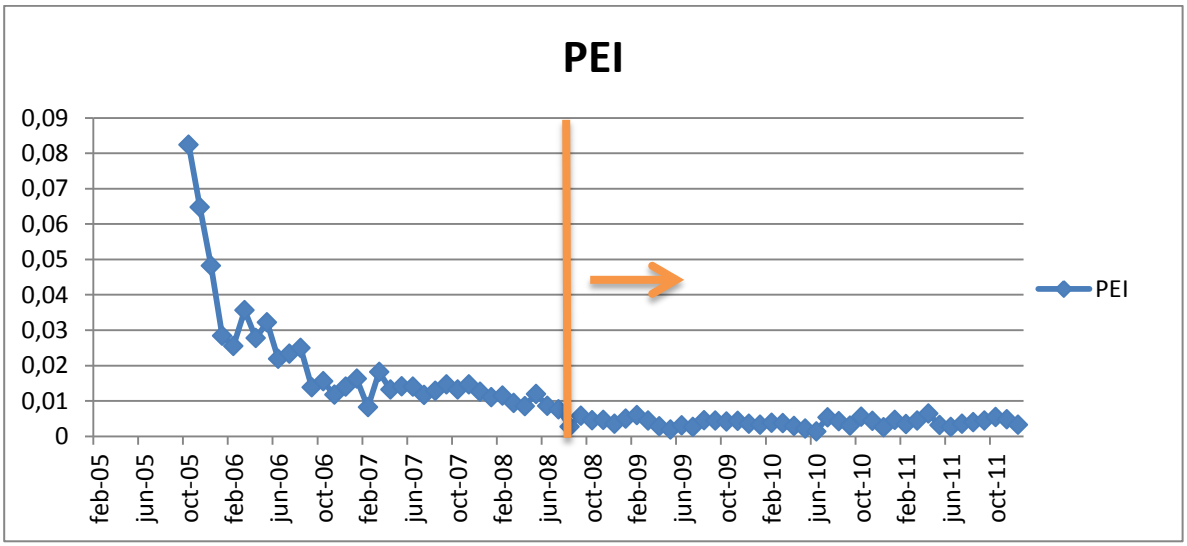

Ilustración 48. Gráfico de la variable PEI a través del tiempo donde se muestra el rango de datos a considerar para el estudio

Cantidad de datos	41
Media	0,392%
Desviación Estándar.	0,001145066
Total outliers sobre 2σ	2

Tabla 37. Análisis de outliers de los datos de la variable PEI

En esta ocasión existen 2 datos outliers correspondientes a Junio 2010 y Abril 2011. Así que se obtuvo la distribución de la variable aleatoria sin considerar esos dos datos mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	Lognormal(0,00393; 0,00109)
p-valor test Chi-Cuadrado	<0,005
p-valor test Kolmogorov-Smirnov	>0,15

Tabla 38. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PEI

Distribución	Error Cuadrático
Lognormal	0,000658
Erlang	0,00109
Gamma	0,00129
Beta	0,00279
Normal	0,00623
Weibull	0,00664
Triangular	0,112
Uniform	0,251
Exponential	0,29

Tabla 39. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PEI

Si bien el p-valor del test Kolmogorov-Smirnov es mayor a 0,15 para la distribución Lognormal, el p-valor del test Chi-Cuadrado es menor a 0,005 y también Lognormal es la distribución con menor error cuadrático.

B.7 Distribución aleatoria de la variable PECt

Este gráfico presenta una estabilidad visual de los datos a partir del mes de Agosto del 2008 en adelante, por lo que consideraremos los datos desde esa fecha.

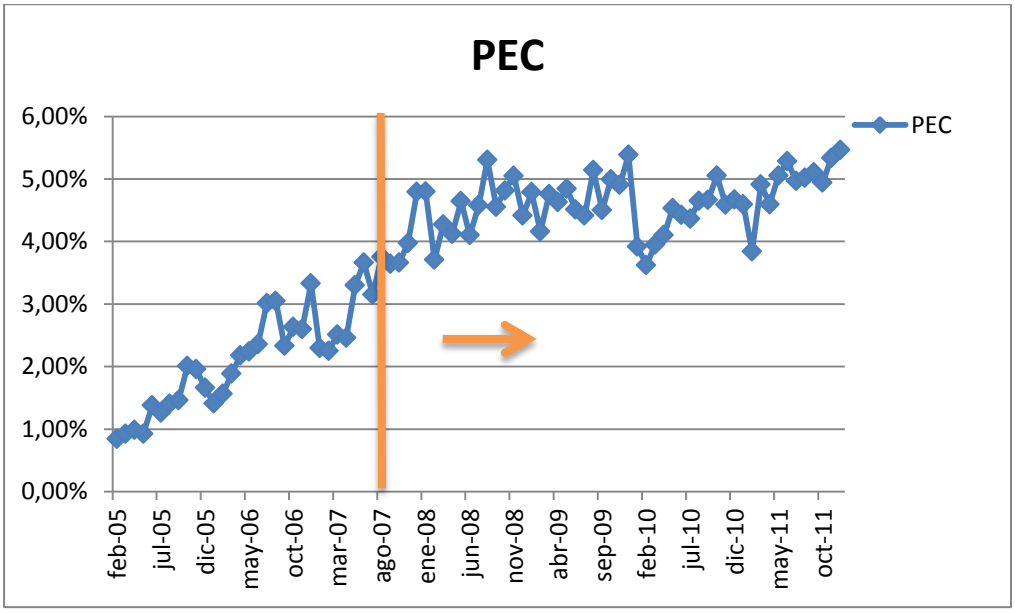

Ilustración 49. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PITC

Cantidad de datos	43
Media	4,69%
Desviación Estándar.	0,435%
Total Outliers sobre 2σ	1

Tabla 40. Análisis de outliers de los datos de la variable PEC

En esta ocasión se tiene 1 outlier correspondiente al dato de Febrero 2010. Así que se obtuvo la distribución de la variable aleatoria sin considerar ese dato mediante la herramienta Input Analyzer del software de Simulación Arena.

Los resultados fueron los siguientes:

Mejor distribución obtenida	0,03 + Weibull(0,0187; 4,92)
p-valor test Chi-Cuadrado	<0,005
p-valor test Kolmogorov-Smirnov	>0,15

Tabla 41. Resultados de Tests Chi-Cuadrado y Kolmogorov Smirnov a los datos de la variable PEC

Distribución	Error Cuadrático
Weibull	0,00903
Beta	0,0142
Normal	0,0146
Gamma	0,0314
Erlang	0,0319
Lognormal	0,0428
Triangular	0,0652
Uniform	0,185
Exponential	0,271

Tabla 42. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable PEC

Si bien el p-valor del test Kolmogorov-Smirnov es mayor a 0,15 para la distribución Weibull, el p-valor del test Chi-Cuadrado es menor a 0,005 y también Weibull es la distribución con menor error cuadrático.

B.8 Distribución aleatoria de la variable MIt

Se observan que los datos varían entre 0 a 4 muertes mensuales. Aemás al calcular el promedio se ve que este corresponde a aproximadamente 1,2 muertes mensuales de imputados dentro de la región metropolitana. Por lo que convendría trabajar un promedio de 1 muerte mensual.

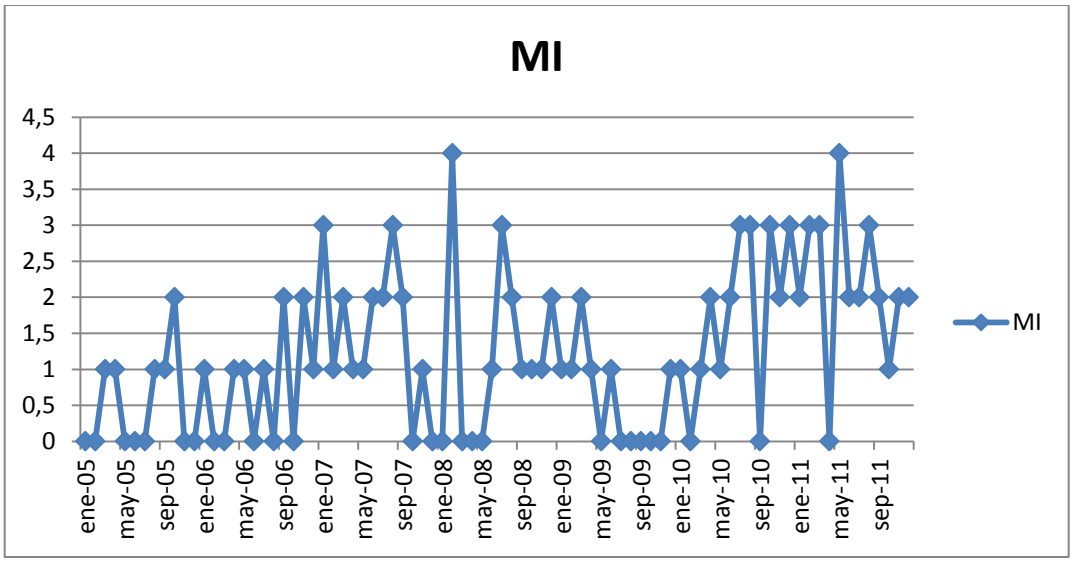

Ilustración 50. Gráfico de la variable MI en cada mes

B.9 Distribución aleatoria de la variable MCt

Aquí los datos se comportan de manera estable a excepción del valor de diciembre 2010 dado que claramente es un outlier. Y además el 8 de Diciembre es la fecha del incendio de la cárcel de San Miguel, lo cual explica el porqué de esa cifra).

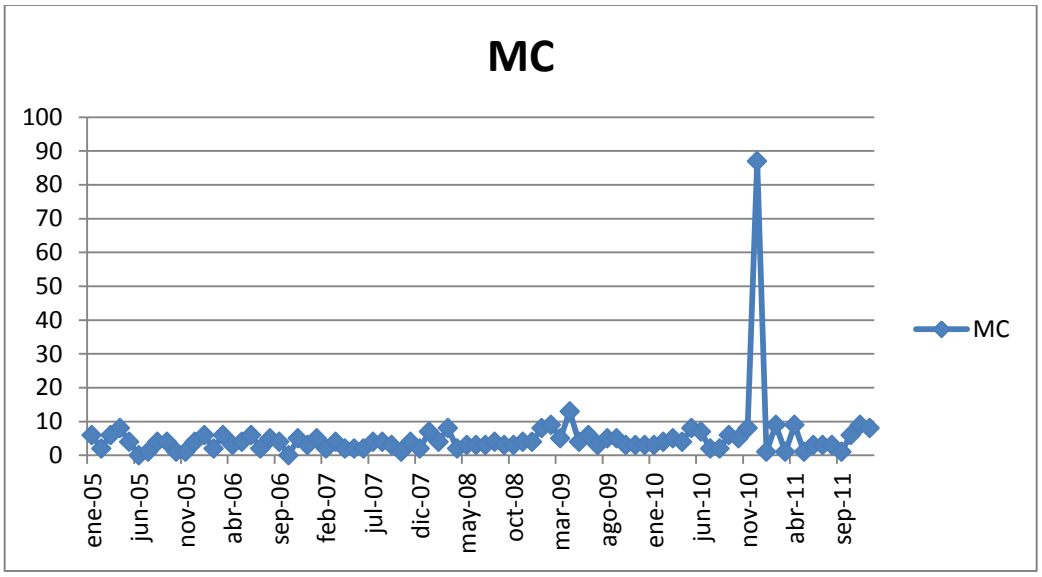

Ilustración 51. Gráfico de la variable MC a través del tiempo donde se muestra el rango de datos a considerar para el estudio

Cantidad de datos	84
Media	5,143
Desviación Estándar.	9,364
Total Outliers sobre 2σ	1

Tabla 43. Análisis de outliers de los datos de la variable MC considerando el dato outlier del incendio de la cárcel de San Miguel

En esta ocasión se ve que hay 1 outlier para un intervalo de ±2σ que corresponde al dato indicado anteriormente a Diciembre del 2010.

Este es un outlier muy grande en comparación con los otros datos (de hecho es casi 17 veces mayor que la media y hace que el resto de los datos se encuentren en un rango casi menor a ±0,5σ.

Ahora si hacemos los mismos cálculos sin considerar ese outlier nos queda lo siguiente.

Cantidad de datos	83
Media	4,157
Desviación Estándar.	2,592
Total Outliers sobre 2σ	1

Tabla 44. Análisis de outliers de los datos de la variable MC sin considerar el dato outlier del incendio de la cárcel de San Miguel

Este outlier corresponde al dato de Abril del 2009 pero es un outlier mucho menor que el de Diciembre del 2010.

A continuación se procederá a obtener la distribución de la variable aleatoria sin considerar esos 2 datos outliers mediante la herramienta Input Analyzer del software de Simulación Arena.

Resultados de Input Analyzer:

Mejor distribución obtenida	-0,5 + ERLA(1,14; 4)
p-valor test Chi-Cuadrado	0,372

Tabla 45. Resultados del Test Chi-Cuadrado a los datos de la variable MC

Distribución	Error Cuadrático
Erlang	0,00728
Gamma	0,00754
Weibull	0,00808
Triangular	0,00831
Poisson	0,00911
Lognormal	0,0121
Normal	0,0129
Beta	0,0145
Uniform	0,0365
Exponential	0,0634

Tabla 46. Resultados del error Cuadrático obtenido para cada distribución probabilística al ajustarse a los datos de la variable MC

En este caso para la distribución Erlang el p-valor del test Chi-Cuadrado es menor a 0,372. Es un p-valor muy alto como para aceptar esta distribución a priori. Pero también Erlang es la distribución con menor error cuadrático.

B.10 Resumen de distribuciones probabilísticas de las variables aleatorias

A continuación se muestra una tabla resumen de las distribuciones probabilísticas de las variables aleatorias del modelo de flujos a simular.

Variable	Distribución
PITC	Weibull(Alpha 2,83; Beta 0,00292)
PITI	Weibull(Alpha 2,39; Beta 0,0027)
PETC	Beta (Alpha1 = 17,4; Alpha 2 = 20,7873)
PETI	Lognormal(Mean =0,0087; Std. Dev. = 0,00283)
CE	74,5 + 75*Beta(Alpha1 = 0,857; Alpha2 = 0,957)
PEI	Lognormal(Mean = 0,00393; Std. Dev. = 0,00109)
PEC	0,03 + Weibull(Alpha = 4,92; Beta = 0,0187)
MI	1 Mensual
MC	Erlang(Mean = 1,14; k = 4) -0,5

Tabla 47. Tabla resumen de las distribuciones probabilísticas de las variables aleatorias