R E V I S T A

INGENIERÍA DE SISTEMAS

Volumen XXV Septiembre 2011

Publicada por el
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

R E V I S T A

INGENIERÍA DE SISTEMAS

ISSN 0716 - 1174

Financiado parcialmente por el Instituto Sistemas Complejos de Ingeniería, como reconocimiento a la difusión de las materias abordadas y de sus participantes.

Instrucciones a los autores:
Los autores deben enviar copia electrónica del manuscrito que desean someter a referato a gduran@dii.uchile.cl. Los manuscritos deben estar escritos a doble espacio, deben incluir un resumen de no más de 150 palabras, y su extensión no debe exceder las 25 páginas. Detalles en www.dii.uchile.cl/∼ris

Correo electrónico: ris@dii.uchile.cl
Web URL: www.dii.uchile.cl/∼ris

Representante legal: Alejandra Mizala
Dirección: República 701, Santiago, Chile.
Diagramación: Cinthya Vergara
Impresión: Ka2 Diseño e Impresión — contacto@ka2.cl
Imagen de Portada: Magnolia Gráfica — www.magnoliagrafica.com

Carta Editorial Volumen XXV

Nos es muy grato presentar este nuevo número de la Revista de Ingeniería de Sistemas (RIS) dedicado a temas de frontera en Investigación de Operaciones, Gestión y Tecnología. Queremos agradecer al Instituto Sistemas Complejos de Ingeniería (ISCI) por su colaboración para hacer posible esta publicación.

Este número contiene artículos de académicos y estudiantes de nuestro Departamento de Ingeniería Industrial (algunos de ellos incluso son consecuencia de trabajos finales de grado, tesis de magister o tesis de doctorado), y de investigadores del ISCI. Contamos también en este número con trabajos de académicos de la Universidad de Concepción y de la Universidad de Buenos Aires, en Argentina.

Nuestro objetivo a través de esta publicación es contribuir a la generación y difusión de las tecnologías modernas de gestión y administración. La revista pretende destacar la importancia de generar conocimiento en estas áreas, orientado tanto a problemáticas nacionales como a la realidad de países de características similares.

Estamos seguros de que los artículos publicados en esta oportunidad muestran formas de trabajo innovadoras que serán de gran utilidad e inspiración para todos los lectores, ya sean académicos o profesionales, por lo que esperamos que esta iniciativa tenga la recepción que creemos se merece.

Guillermo Durán
Editor

El Fenómeno del Plagio en Documentos Digitales: Un Análisis de la Situación Actual en el Sistema Educacional Chileno

Francisco Molina, Juan D. Velásquez,Sebastián Ríos [*]
Paulina A. Calfucoy, Matías Cociña V. [*]

Resumen

Este artículo presenta el estado actual del plagio digital en la educación chilena, mediante los resultados de la "Encuesta acerca de Prácticas de Estudio y Trabajo de Estudiantes de Educación Media y Superior", diseñada y ejecutada en el marco del Proyecto Fondef «DOcument COpy DEtector»(DOCODE). El principal objetivo de la encuesta fue identificar las percepciones y prácticas de alumnos de educación media y superior en torno al plagio. Específicamente, se intentó describir cómo los estudiantes entienden el plagio, cuáles son sus prácticas, bajo qué condiciones plagian y cómo justifican esta práctica. La encuesta fue aplicada en las regiones Metropolitana y de Valparaíso, y consideró 2.031 alumnos de Educación Media (1.397 de la RM y 634 de la V Región) y 1.126 estudiantes de Educación Superior (todos de la RM). Por otro lado, se abordaron 5 temas principales: i) características socioeconómicas y socio-demográficas; ii) acceso y pautas de consumo de Internet y tecnologías de información; iii) conocimiento acerca de buenas prácticas para el desarrollo de trabajos escritos; iv) prácticas de plagio de los estudiantes y percepciones respecto de las prácticas de sus pares; y, finalmente, v) percepción acerca de las potenciales razones que permitirían justificar estas prácticas. Como resultado, se obtuvo que un 55 % de los estudiantes de educación media, y un 42 % de los alumnos de educación superior, declaran haber copiado y pegado información desde la Web sin citar la fuente, lo que parece justificarse por razones de orden pragmático, sustentadas, en cierta forma,

[*]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

por la falta de conocimiento respecto de la noción de plagio, y facilitadas por una masificación del acceso a computadores e Internet.

Palabras Clave: Plagio Digital, Copy/Paste, Educación.

1. Introducción

Actualmente, la sociedad enfrenta una transformación importante en la manera en que gestiona el conocimiento. El acceso a las redes digitales y la disponibilidad de enormes cantidades de información han adquirido una centralidad nunca antes vista [3]. En este contexto, la masificación de internet no sólo ha cambiado considerablemente la forma en que los estudiantes acceden a la información para el desarrollo de sus experiencias educativas [16], sino que también ha impactado en las prácticas de los profesores, quienes enfrentan hoy nuevos desafíos como la evaluación de la originalidad de las ideas de sus alumnos, y el potencial plagio de trabajos escritos utilizando información disponible en la Web [4, 5, 12]. En respuesta a lo anterior, se ha generado una creciente preocupación por parte de docentes y profesores por comprender los cambios en las prácticas de los estudiantes al momento de hacer sus trabajos escritos, surgiendo la necesidad de desarrollar instrumentos que permitan facilitar la detección de plagio digital [15].

Para comprender por qué y bajo qué condiciones los estudiantes plagian, diversos autores han expuesto diferentes explicaciones alternativas. Entre ellas se encuentran: presión o apuro, falta de conocimiento sobre las *reglas del juego* y lo difícil que resulta detectar el "cyberplagio" para los profesores [6]. Por otro lado, existen visiones más pragmáticas del problema que afirman que los alumnos plagian, simplemente, porque pueden hacerlo [14].

De acuerdo a la revisión de literatura realizada, en [15] se ha demostrado que existe una estrecha relación entre diferentes variables del contexto y la disposición de los estudiantes a copiar, entre las cuales se incluyen la percepción acerca del comportamiento de los pares, la comprensión y aceptación de políticas de ética, la percepción acerca de la probabilidad de ser descubierto copiando y la percepción acerca de la severidad de las penas. Dentro de estas variables del contexto, la percepción acerca del comportamiento de los pares pareciera ser la más importante para entender la incidencia de plagio entre los estudiantes [7, 8, 9].

Algunas cifras del plagio académico en el mundo

Un estudio realizado el año 2001 por McCabe [11], el cual abarcó 2.294 estudiantes de secundaria de 25 escuelas públicas y privadas de EE.UU., entregó que un 52 % de los alumnos declaró haber copiado párrafos explícitos desde algún sitio web sin realizar la correspondiente cita.

Del mismo modo, una encuesta realizada por el Josephson Institute of Ethics [13], realizada a más de 36.000 estudiantes de secundaria, entregó niveles alarmantes de plagio ya que el 60 % de los alumnos admitió haber cometido engaño mediante la copia de documentos dentro de los últimos 12 meses, mientras que uno de cada tres estudiantes reconoció haber utilizado la web para plagiar información en alguno de sus trabajos.

Por otro lado, cabe destacar una investigación sudamericana realizada por Bordigon et al. [1], la cual señaló que el 50 % de los alumnos de educación básica y media del medio escolar argentino declaran haber copiado para confeccionar sus tareas y trabajos.

El plagio académico en el Chile

Para profundizar en el fenómeno del plagio digital en la educación chilena, a continuación se exponen los resultados de un estudio realizado en el marco del Proyecto Fondef DOCODE [1], el cual pretende dar respuesta a varias de estas interrogantes, enfocándose en la educación media y superior de Chile.

El documento comienza con la sección 2, la cual contiene las características de la encuesta realizada (diseño del instrumento, diseño muestral y aplicación del instrumento). Luego, en la sección 3, se exponen los resultados obtenidos, centrando el análisis en las percepciones y prácticas de plagio de los estudiantes y sus pares. Finalmente, en la sección 4, se presentan las conclusiones con los principales aprendizajes logrados a partir de los resultados de la encuesta.

2. Diseño del instrumento, diseño muestral y aplicación del instrumento

El diseño del instrumento de medición se organizó en torno a cinco conceptos: (a) caracterización del estudiante; (b) desempeño académico; (c) alfabetización digital; (d) prácticas al hacer tareas y trabajos; y (e) plagio. Cada uno de estos conceptos se tradujo en una o más dimensiones, que luego se operacionalizaron en variables, y éstas a su vez en preguntas a incorporar en el instrumento.

La *caracterización del estudiante* se operacionalizó a través de sus rasgos

[1] Para obtener mayor información del proyecto, visitar el sitio web `http://www.docode.cl`

socio-demográficos; el *desempeño escolar* mediante el desempeño escolar auto-reportado; la *alfabetización digital* a través del acceso de los estudiantes a tecnologías de la información y el propósito de estos en el uso de internet; las *prácticas para hacer tareas* mediante el manejo de fuentes de información y el uso de internet como fuente de información, además del conocimiento de buenas prácticas en el trabajo con fuentes y la claridad en las instrucciones recibidas por el profesor respecto del plagio. Finalmente, en la sección *plagio* se estudió la definición de plagio y su entendimiento por parte de los alumnos, las prácticas individuales de plagio, las percepciones respecto de las prácticas de otros y las justificaciones esgrimidas por los alumnos para, eventualmente, plagiar textos.

El instrumento fue diseñado entre abril y marzo de 2010, realizándose el trabajo de campo entre mayo y agosto del mismo año.

3. Situación en Chile, Presentación de Resultados

La siguiente sección da cuenta de los principales resultados de la encuesta realizada. Con esta información se busca mejorar la comprensión de las prácticas de plagio entre estudiantes chilenos y las razones y circunstancias que afectan su incidencia.

3.1. Resultados muestra educación secundaria

3.1.1. Descripción sociodemográfica de la muestra

El 55 % de la muestra de alumnos de colegios corresponde a hombres, mientras que el 45 % corresponde a mujeres. El muestreo representa de manera relativamente homogénea todos los niveles de educación media (I a IV medio), sobre representando levemente a los alumnos de II medio (31 %). La mediana y la media de edad en la muestra son 16 años, con edades mínima y máxima de 12 y 21 años de edad, respectivamente, y con una desviación estándar de 1,4 años. El 95 % de los alumnos en la muestra tienen entre 14 y 18 años de edad.

La distribución de notas (promedio el año anterior, reportado por los alumnos) se muestra en la Tabla 1:

Adicionalmente se pidió a los estudiantes informar el nivel educacional del jefe de hogar. La Tabla 2 muestra la distribución de nivel educacional del jefe de hogar, condicional al nivel socioeconómico del colegio o escuela. La distribución de nivel educacional por nivel socioeconómico de la institución es consistente

Rango Notas	Porcentaje
[6.0 - 7.0]	18,7
[5.0 - 5.9]	61,4
[4.0 - 4.9]	17,6
[1.0 - 3.9]	2,3

Tabla 1: Promedio notas año anterior (Ed. Secundaria)

la distribución esperable (mayores niveles educacionales en instituciones de mayor nivel socioeconómico).

Nivel Educacional del Jefe de Hogar	Alto	Medio-Alto	Medio	Medio-Bajo	Bajo	Total
Escolar Incompleta	3,8	25,4	31,3	53,0	56,2	38,3
Secundaria Completa	5,3	27,5	38,4	32,7	25,3	29,9
Técnico-Prof. Incompleta	1,8	5,2	2,2	1,4	0,8	2,2
Técnico-Prof. Completa	10,0	17,2	14,5	6,6	7,3	10,7
Universitaria Incompleta	4,5	2,6	2,4	1,3	2,4	2,2
Universitaria Completa	74,5	22,1	11,3	4,9	7,8	16,7
Total	100,0	100,0	100,0	100,0	100,0	100,0

Tabla 2: Nivel Educacional del Jefe de Hogar (agrupado), por Nivel Socioeconómico del Colegio

Todo lo anterior sugiere que la muestra de estudiantes de educación secundaria no presenta anomalías respecto de las variables socioeconómicas y demográficas. El proceso de selección del número de alumnos encuestados por estrato (Tipo de dependencia del colegio - Nivel socioeconómico del colegio) se detalla en el Anexo A.

3.1.2. Análisis descriptivo de los resultados

La siguiente sección muestra los principales resultados obtenidos respecto del nivel de acceso que los estudiantes tienen a recursos computacionales, el uso que hacen de ellos, y cómo los utilizan en su interacción con la escuela o colegio. Finalmente, se presentan resultados descriptivos sobre actitudes, percepciones y prácticas respecto del plagio en trabajos escritos.

Acceso y uso del computador e internet

Más del 80 % de los estudiantes encuestados declaran tener acceso a compu-

tador e internet como herramienta de trabajo en el marco de su trabajo escolar. Más aún, más de un 65 % de los estudiantes tiene acceso permanente al computador como herramienta de trabajo, y a internet como fuente de información. La proporción de alumnos que declaran no tener acceso a estos recursos es baja.

Respecto del uso que los estudiantes dan a dichos recursos, destaca la alta frecuencia con que los alumnos utilizan internet para comunicarse con sus pares: más del 50 % de los alumnos la utilizan para estos fines al menos cinco días a la semana, mientras que un 35 % lo hace todos los días (la mediana son 4 días). El uso de internet como medio de entretenimiento (la mediana son 3 días) tiene la segunda proporción más alta de alumnos que realizan esta actividad todos los días, pero concentra un alto porcentaje de alumnos que juegan dos, uno, o ningún día a la semana. La búsqueda de información y la realización de tareas y trabajos, pese a tener medianas similares, tiene menos valores extremos: una baja proporción (16.7 %) de estudiantes busca información o hace tareas o trabajos todos los días, y una proporción aún más baja no lo hace nunca.

Prácticas de trabajo en la realización de tareas y trabajos escritos

Dado el acceso prácticamente universal, y una frecuencia de uso de internet para la realización de trabajos relativamente alta (tres días en promedio), es posible esperar que internet sea una fuente relevante de información para la realización de tareas y trabajos en el marco de las actividades escolares. Al estudiar las fuentes consultadas por los alumnos para la realización de sus trabajos, internet destaca como la principal fuente de información para la realización de trabajos escritos por parte de los alumnos de educación secundaria: más de un 75 % de los alumnos dice utilizar internet "siempre" o "casi siempre" como fuente de información. El contraste con fuentes tradicionalmente utilizadas hasta hace algunos años, como fichas de estudio o revistas, es marcado. Menos de un 30 % de los alumnos declara utilizar libros como fuente de información "siempre" o "casi siempre", y más de un tercio de los alumnos dice utilizar libros "casi nunca" o "nunca".

Fuentes consultadas en Internet

Según los datos analizados, más del 85 % de los alumnos utiliza "siempre" o "casi siempre" buscadores como su principal fuente de información (Google, Yahoo!, etc.). Luego, un 68 % ocupa, como su segunda fuente de información, sitios como Wikipedia u otra enciclopedia en línea. Por otro lado, los sitios menos utilizados son los de contenido educativo complementario, como el sitio de la tradicional revista Icarito o educarchile.cl (sólo el 33 % de los alumnos los utilizan "siempre" o "casi siempre"); y sitios de tareas y trabajos ya escritos, como El Rincón del Vago (sólo el 23 % de los alumnos los utiliza "siempre" o

"casi siempre").

Plagio: definiciones y conocimiento por parte de los alumnos

Para evaluar el nivel de conocimiento respecto del concepto y práctica del plagio, se presentó a los alumnos cuatro definiciones alternativas de plagio y se les solicitó determinar si cada definición correspondía o no, a su juicio, a una acción de "plagio o copia". En la tabla 3 se presentan los resultados ante la afirmación principal "Creo que es copia o plagio buscar información desde Internet, desde un libro o una revista y...".

	Si es copia/plagio	No es copia/plagio	Total
...usar esta información en un trabajo sin cambiar el contenido y sin citar desde donde se sacó la información	82,4 %	17,6 %	100 %
...usar la información para redactar con las propias palabras citando la fuente	21,6 %	78,4 %	100 %
...cambiar algunas de las palabras manteniendo la idea original, pero sin citar desde donde saqué la información	48,8 %	51,2 %	100 %
...usar las ideas de otros como mías sin hacer referencia al autor	81,3 %	18,7 %	100 %

Tabla 3: De las siguientes maneras de hacer trabajos escritos, ¿cuáles consideras que son copia o plagio?

De las cuatro definiciones, la primera y la cuarta son definiciones absolutas de plagio o copia, que en un principio no se debiesen prestar a confusión. El resultado es consistente: más del 80 % de los encuestados identifica correctamente estos casos como plagio. Aún así, es interesante constatar que hay casi un 20 % de los estudiantes de educación secundaria que no logran distinguir una práctica de plagio siquiera en su versión más absoluta.

La segunda definición ofrecida ("usar información para redactar con palabras propias, citando la fuente") es evidentemente un caso que no corresponde a plagio. Los alumnos, nuevamente, en su gran mayoría identifican correc-

tamente el caso. Sin embargo, una vez más, hay cerca de un quinto de los alumnos que no logran establecer esta diferencia.

Finalmente, la tercera definición, que también corresponde a un caso de plagio, es identificada como tal por menos de la mitad de los encuestados. Un punto a destacar es que esta definición, algo más confusa, probablemente describe en buena medida la actividad de copia por parte de los alumnos.

Por otro lado, si bien el porcentaje de alumnos que ha recibido formación respecto de cómo citar no es despreciable (casi un 50 %), existe una brecha de aprendizaje entre la búsqueda de información, el conocimiento respecto del carácter inadecuado de la copia, y el conocimiento de prácticas de uso de información que eviten la copia. Los alumnos parecen saber cómo buscar información y conocer del carácter negativo del plagio, pero al parecer carecen de las herramientas que les permitan citar las fuentes utilizadas de manera adecuada.

El plagio y sus justificaciones

El porcentaje de alumnos que declara haber incurrido en prácticas de plagio "más de una vez" o "muchas veces" (durante el año anterior) supera el 55 %. La misma proporción de alumnos cree que "Todos" o "Casi todos" sus compañeros copiaron al menos una vez en el año previo a la encuesta. Por otro lado, el 80 % de alumnos que declara haber copiado y pegado información de internet sin citar la fuente, al menos una vez el año anterior.

Para explorar las razones que los alumnos consideran válidas para justificar un comportamiento de plagio, se utilizó un serie de razones mencionadas en la literatura sobre plagio: razones utilitarias; de (falta de) confianza personal; ignorancia; y razones adaptativas.

Como se puede apreciar en la tabla 4, el ahorro de tiempo (utilitaria), así como la creencia de que el trabajo queda mejor si es plagiado (confianza personal), son las razones más validadas por los alumnos.

A continuación (aunque a considerable distancia), está la justificación adaptativa de falta de tiempo donde el plagio es una respuesta a una imposición de restricciones exógenas. Por otro lado, la ignorancia y otras razones adaptativas son consideradas las menos válidas para justificar el plagio.

En suma, el plagio es una práctica extendida a nivel de los alumnos de educación secundaria. En parte, justificada por razones de orden pragmático que se sustentan, en alguna medida, en una falta de conocimiento respecto de la noción de plagio y su aplicabilidad. La práctica se ve facilitada por una masificación del acceso a los computadores e Internet, que ha llegado a ser la principal fuente de información de los alumnos. Con todo esto, se cree que el espacio de aprendizaje entre profesores y alumnos tiene el potencial de corregir los efectos de la ignorancia respecto del concepto de plagio.

	Muy Mala	Mala	Regular	Buena	Muy Buena	Total
Permite ahorrar tiempo	6,8	10,3	28,8	30,0	24,2	100
El trabajo queda mejor que si lo hago por mi mismo	9,7	15,3	22,7	19,0	33,4	100
Me es difícil escribir o sintetizar información	16,2	24,7	35,0	14,5	9,6	100
No sé hacerlo de otra manera	37,4	31,6	19,6	6,3	5,0	100
Es lo que me piden	27,0	22,4	27,8	14,2	8,6	100
A la gente que copia le va mejor	34,6	27,7	21,0	8,4	8,3	100
Lo hago porque todos lo hacen	40,6	29,3	16,7	7,6	5,8	100
Nunca creí que era un problema	21,9	27,2	31,2	10,4	9,4	100
Creo que al profesor no le importa	27,8	27,5	24,0	11,3	9,4	100
Creo que el profesor no se da cuenta	25,6	29,2	24,2	10,0	11,0	100
No me queda tiempo para hacerlo de otra manera	19,0	21,6	27,9	15,8	15,6	100

Tabla 4: Califica en la siguiente escala, ¿en qué medida las siguientes razones para copiar y pegar sin citar representan una buena justificación para ti?

Cruce de variables

Para estudiar qué variables pueden afectar (o correlacionarse con) las prácticas de plagio de los alumnos, se utilizó el porcentaje de alumnos que copiaron "más de una vez" o "muchas veces" el año anterior a la encuesta, versus aquellos que copiaron "una vez" o "nunca". Luego, se analizó la variación de dichos porcentajes condicional a otras variables.

En primer lugar, cabe mencionar que las prácticas de plagio no varían

significativamente de acuerdo al sexo de los alumnos. Además, el nivel socioeconómico del colegio, así como el nivel educacional de los padres, tampoco influye en los porcentajes de copia, aunque el porcentaje de copia en el nivel socioeconómico alto es levemente mayor. Por otro lado, el porcentaje de plagio sí varía de acuerdo al tipo de dependencia administrativa del establecimiento: el nivel de plagio es significativamente mayor en colegios particulares pagados que en subvencionados, y en subvencionados que en municipales. Esta diferencia puede ser explicada, al menos en parte, por el diferencial de acceso a computadores e Internet entre estos tres grupos. (Ver Tabla 5.)

Dependencia administrativa del establecimiento	Una vez o nunca	Más de una vez (Más de una o Muchas veces)	Total
Municipal	51,8	48,2	100,0
Particular Subvencionado	42,9	57,1	100,0
Particular pagado	38,2	61,8	100,0
Total	44,5	55,5	100,0

Tabla 5: Dependencia administrativa vs. Frecuencia de copia

Otra diferencia significativa es que el nivel de plagio es menor entre los alumnos que presentan mejores resultados académicos; mientras que en el grupo con promedio entre 4.0 y 4.9 los alumnos con mayor tasa de plagio representan un 62 %, en el grupo con notas sobre 6.0 la cifra cae a 52 %.

Por otro lado, la correlación más evidente con la frecuencia de plagio viene dada por las creencias de los alumnos respecto del comportamiento de sus pares. Entre aquellos estudiantes que creen que sólo algunos o ninguno de sus compañeros plagia sus trabajos, la proporción de alumnos que plagiaron reiteradamente es menor al 25 %. En cambio, entre aquellos que creen que casi todos o todos sus compañeros plagiaron al menos una vez el año anterior, más de un 70 % plagió repetidamente el año anterior. (Ver Tabla 6.)

Respecto de las prácticas de trabajo de los alumnos, el hecho de que éstos consulten con profesores o con compañeros para la realización de sus trabajos, parece no incidir en su propensión al plagio. Tampoco parece incidir con qué frecuencia usan Internet como fuente de información. Por el contrario, aquellos alumnos que consultan libros más frecuentemente, y aquellos que consultan con sus padres, presentan menores tasas de plagio. Así, por ejemplo, entre aquellos que nunca consultan libros, un 63 % declara haber plagiado más de una vez, mientras que entre aquellos alumnos que consultan libros siempre, dicho porcentaje es de un 35 %.

¿Qué porción de tus compañeros de curso crees que copiaron y pegaron sin citar la fuente al menos una vez en un trabajo?	Una vez o nunca	Más de una vez (Más de una o Muchas veces)	Total
Ninguno	80,1	19,9	100,0
Algunos	77,6	22,4	100,0
La mitad	58,0	42,0	100,0
Casi todos	28,9	71,1	100,0
Todos	14,0	86,0	100,0
Total	44,5	55,5	100,0

Tabla 6: Cantidad de compañeros que copiaron y pegaron sin citar en un trabajo vs. Frecuencia de copia

Por otro lado, es interesante mencionar que las percepciones de qué constituye plagio no parece cambiar ente aquellos que copian más frecuentemente y aquellos que no lo hacen (la variación es siempre menor a 4 puntos porcentuales). En el agregado, la diferencia en el nivel de plagio no parece venir dada por diferenciales en la comprensión de qué constituye plagio. Sin embargo, las justificaciones que más validez tienen entre los estudiantes que más plagian, y la que más distingue a este grupo respecto de aquellos estudiantes que no realizan plagio, es el ahorro de tiempo o la escasez de tiempo. (Ver tabla 7.)

3.2. Resultados muestra educación superior

3.2.1. Descripción sociodemográfica de la muestra

El 48.4 % de alumnos de educación superior corresponde a hombres, mientras que el 51.7 % a mujeres. El muestreo representa de manera homogénea los años Primero a Quinto (o más). La mediana es 22 años y la media es 23.4 años, con edad mínima de 17 y máxima de 58 años, y con una desviación estándar de 5.6 años. Más del 80 % de la muestra se concentra entre los 18 y 25 años.

La distribución de las notas (promedio año anterior) se puede apreciar en la Tabla 8.

Adicionalmente, se pidió a los estudiantes informar el nivel educacional del jefe de hogar. La distribución de la variable, agrupada en seis categorías, se muestra en la Tabla 9.

Todo lo anterior sugiere que la muestra de estudiantes de educación secundaria no presenta grandes sesgos en las variables medidas.

Justificación:	Una vez o nunca	Más de una vez (Más de una o Muchas veces)	Total
Permite ahorrar tiempo	40,8	65,0	54,2
El trabajo queda mejor que si lo hago por mi mismo	52,7	52,1	52,33
Me es difícil escribir o sintetizar información	19,0	28,2	24,1
No sé hacerlo de otra manera	10,2	12,3	11,3
Es lo que me piden	22,9	22,8	22,8
A la gente que copia le va mejor	13,8	19,1	16,7
Lo hago porque todos lo hacen	10,1	16,2	13,5
Nunca creí que era un problema	13,7	24,6	19,8
Creo que al profesor no le importa	14,3	25,9	20,7
Creo que el profesor no se da cuenta	14,9	26,0	21,1
No me queda tiempo para hacerlo de otra manera	23,4	37,9	31,5

Tabla 7: Porcentaje de alumnos que considera la justificación (para el plagio) "Buena" o "Muy Buena" vs. Frecuencia de copia

Rango Notas	Porcentaje
[6.0 - 7.0]	12,3
[5.0 - 5.9]	64,5
[4.0 - 4.9]	22,5
[1.0 - 3.9]	0,7

Tabla 8: Promedio notas año anterior (Ed. Superior)

3.2.2. Análisis descriptivo de los resultados

La siguiente sección muestra los principales resultados obtenidos respecto del nivel de acceso que los estudiantes tienen a recursos computacionales, el uso que hacen de ellos, y cómo los utilizan en su interacción con las institu-

	Total
Escolar Incompleta	15,9
Secundaria Completa	25,5
Técnico-Profesional Incompleta	4,8
Técnico-Profesional Completa	17,3
Universitaria Incompleta	7,6
Universitaria Completa	29,0
Total	100,0

Tabla 9: Nivel Educacional del Jefe de Hogar (agrupado)

ciones donde estudian. Finalmente, se presentan resultados descriptivos sobre actitudes, percepciones y prácticas respecto del plagio en trabajos escritos.

Acceso y uso del computador e internet

Más del 95 % de los estudiantes encuestados declaran tener acceso a computador e internet como herramienta de trabajo en el marco de su trabajo escolar. Más aún, más de un 79 % de los estudiantes tiene acceso permanente al computador como herramienta de trabajo, y a Internet como fuente de información. La proporción de alumnos que declaran no tener acceso a estos recursos es cercana a cero.

Respecto del uso que los estudiantes dan a dichos recursos, destaca la alta frecuencia con que los alumnos utilizan Internet para comunicarse con sus pares: más del 70 % de los alumnos chatean o acceden a sitios sociales al menos cinco días a la semana, mientras que un 49 % lo hace todos los días (mediana = 6 días). A diferencia del caso de los colegios, el uso de Internet como medio de entretenimiento (mediana = 4) es la menos frecuente de las actividades en línea que realizan los alumnos. La búsqueda de información y la realización de trabajos presentan patrones muy similares (mediana = 5).

Prácticas de trabajo en la realización de tareas y trabajos escritos

Al estudiar las fuentes consultadas por los alumnos para la realización de sus trabajos, internet destaca una vez más como la principal fuente de información para la realización de trabajos escritos por parte de los alumnos de educación secundaria: casi un 95 % de los alumnos dice utilizar Internet "siempre" o "casi siempre" como fuente de información. Comparado con los estudiantes de educación secundaria, más alumnos (40 %) declaran utilizar libros como fuente de información "siempre" o "casi siempre", y más de un cuarto de los alumnos dice utilizar libros "casi nunca" o "nunca".

Fuentes consultadas en Internet

Según los datos analizados, más del 95 % de los alumnos utiliza "siempre"

o "casi siempre" buscadores como su principal fuente de información (Google, Yahoo!, etc.). Luego, un 55 % ocupa, como su segunda fuente de información, sitios como Wikipedia u otra enciclopedia en línea. Por otro lado, los sitios menos utilizados son los de contenido educativo complementario, como el sitio de la revista Icarito o educarchile.cl (sólo el 22 % de los alumnos los utilizan "siempre" o "casi siempre"); y sitios de tareas y trabajos ya escritos, como El Rincón del Vago (sólo el 14 % de los alumnos los utiliza "siempre" o "casi siempre").

Plagio: definiciones y conocimiento por parte de los alumnos

Para evaluar el nivel de conocimiento respecto del concepto y práctica del plagio, se presentó a los alumnos cuatro definiciones alternativas de plagio y se les solicitó determinar si cada definición correspondía o no, a su juicio, a una acción de "plagio o copia". En la Tabla 10 se presentan los resultados ante la afirmación principal "Creo que es copia o plagio buscar información desde Internet, desde un libro o una revista y...".

	Si es copia/plagio	No es copia/plagio	Total
...usar esta información en un trabajo sin cambiar el contenido y sin citar desde donde se sacó la información	93,6 %	6,4 %	100 %
...usar la información para redactar con las propias palabras citando la fuente	13,5 %	86,5 %	100 %
...cambiar algunas de las palabras manteniendo la idea original, pero sin citar desde donde saqué la información	70,9 %	29,1 %	100 %
...usar las ideas de otros como mías sin hacer referencia al autor	88,2 %	11,8 %	100 %

Tabla 10: De las siguientes maneras de hacer trabajos escritos, ¿cuáles consideras que son copia o plagio?

Por otro lado, el porcentaje de alumnos que han recibido formación respecto de cómo citar es cercano al 50 %, lo que es levemente más alto que en el caso de los estudiantes de educación secundaria. Los alumnos declaran, sin embargo, haber recibido menos formación en cómo buscar información. El nivel de advertencia respecto de la prohibición de plagiar es también más alto.

El plagio y sus justificaciones

El porcentaje de alumnos que declara haber incurrido en prácticas de plagio "más de una vez" o "muchas veces" (durante el año anterior) supera el 42 %. Una proporción similar de alumnos (39 %) cree que "Todos" o "Casi todos" sus compañeros copiaron al menos una vez en el año previo a la encuesta. Por otro lado, el 64 % de alumnos que declara haber copiado y pegado información de internet sin citar la fuente, al menos una vez el año anterior.

Para explorar las razones que los alumnos consideran válidas para justificar el plagio, se utilizó un serie de razones mencionadas en la literatura sobre plagio: utilitarias; de (falta de) confianza personal; ignorancia; y adaptativas.

Como muestra la Tabla 11, el ahorro de tiempo (utilitaria), así como la creencia de que el trabajo queda mejor si es plagiado (confianza personal), son nuevamente las justificaciones más validadas por los alumnos. Sin embargo, estos niveles de justificación son marcadamente menores que entre los estudiantes de educación media.

En suma, el plagio es una práctica extendida a nivel de los alumnos de educación superior. En parte, justificada por razones de orden pragmático que se sustentan, en alguna medida, en una falta de conocimiento respecto de la noción de plagio y su aplicabilidad. La práctica se ve facilitada por una masificación del acceso a los computadores e Internet, que ha llegado a ser la principal fuente de información de los alumnos. Con todo esto, se cree que el espacio de aprendizaje entre profesores y alumnos tiene el potencial de corregir los efectos de la ignorancia respecto del concepto de plagio.

Cruce de variables

Al igual que la muestra de estudiantes secundarios, se utilizó el porcentaje de alumnos que copiaron "Más de una vez" o "Muchas veces" el año anterior a la encuesta, versus aquellos que copiaron "Una vez" o "Nunca". Luego, se analizó cómo varían dichos porcentajes condicional a otras variables.

Nuevamente, las prácticas de plagio no varían significativamente de acuerdo al género. Sin embargo, el nivel educacional del jefe de hogar parece estar correlacionado con la frecuencia de plagio; alumnos con un jefe de hogar de mayor nivel educacional, presentan menor nivel de plagio. (Ver Tabla 12.)

Por otro lado, a diferencia de los alumnos de enseñanza media, el patrón de plagio en alumnos de educación superior no parece estar correlacionado con el nivel de notas del alumno.

Además, nuevamente la correlación más evidente con la frecuencia de plagio, viene dada por la percepción de los alumnos respecto del comportamiento de sus pares: entre aquellos alumnos que creen que sólo "algunos" de sus compañeros plagian sus trabajos, la proporción de alumnos que plagiaron reiteradamente es cercana al 15 %. Sin embargo, entre aquellos que creen que "casi

	Muy Mala	Mala	Regular	Buena	Muy Buena	Total
Permite ahorrar tiempo	13,2	18,2	27,1	25,9	15,6	100
El trabajo queda mejor que si lo hago por mi mismo	24,7	22,4	16,5	14,4	22,0	100
Me es difícil escribir o sintetizar información	24,9	32,1	27,8	11,5	3,8	100
No sé hacerlo de otra manera	58,9	25,3	10,8	3,7	1,3	100
Es lo que me piden	49,0	22,6	16,0	8,1	4,3	100
A la gente que copia le va mejor	53,8	21,6	16,1	5,2	3,4	100
Lo hago porque todos lo hacen	59,2	24,3	11,0	3,9	1,6	100
Nunca creí que era un problema	40,0	26,5	22,7	7,3	3,5	100
Creo que al profesor no le importa	42,9	24,4	20,8	8,8	3,1	100
Creo que el profesor no se da cuenta	43,4	25,2	19,7	8,8	2,8	100
No me queda tiempo para hacerlo de otra manera	30,4	20,7	27,1	14,6	7,2	100

Tabla 11: Califica en la siguiente escala, ¿en qué medida las siguientes razones para copiar y pegar sin citar representan una buena justificación para ti?

todos" o "todos" sus compañeros plagiaron al menos una vez el año anterior, más de un 70 % plagió repetidamente el año anterior. Este patrón es similar al encontrado para la muestra de estudiantes secundarios. (Ver Tabla 13.)

Respecto de las prácticas de trabajo de los alumnos, la correlación entre utilización de libros como fuente de información, y la intensidad de plagio, sigue siendo negativa y significativa. Además, aquellos alumnos que acceden más frecuentemente a profesores para obtener información para sus trabajos,

Nivel educacional del jefe de hogar	Una vez o nunca	Más de una vez (Más de una o Muchas veces)	Total
Escolar incompleta	53,3	46,7	100,0
Secundaria completa	50,1	49,9	100,0
Técnico-Profesional incompleta	55,4	44,6	100,0
Técnico-Profesional completa	58,8	41,2	100,0
Universitaria incompleta	69,5	30,5	100,0
Universitaria completa (con o sin postgrado)	61,5	38,5	100,0

Tabla 12: Nivel educacional del jefe de hogar versus Frecuencia de copia

¿Qué porción de tus compañeros de curso crees que copiaron y pegaron sin citar la fuente al menos una vez en un trabajo?	Una vez o nunca	Más de una vez (o Muchas veces)	Total
Ninguno	100,0	0,0	100,0
Algunos	84,2	15,8	100,0
La mitad	61,1	38,9	100,0
Casi todos	28,9	71,1	100,0
Todos	12,4	87,6	100,0
Total	57,1	42,9	100,0

Tabla 13: Cantidad de compañeros que copiaron y pegaron sin citar en un trabajo versus Frecuencia de copia

tienen menos probabilidades de pertenecer al grupo de plagio más frecuente.

Por otra parte, en términos de las justificaciones más validadas por los alumnos de educación superior, las variables mejor evaluadas son similares a las observadas para educación secundaria: ahorro de tiempo y aumento de calidad en el trabajo. (Ver tabla 14.)

Justificación:	Una vez o nunca	Más de una vez (o Muchas veces)	Total
Permite ahorrar tiempo	31,4	55,3	41,6
El trabajo queda mejor que si lo hago por mi mismo	32,3	41,8	36,4
Me es difícil escribir o sintetizar información	10,2	22,2	15,3
No sé hacerlo de otra manera	4,2	6,2	5,0
Es lo que me piden	10,4	15,2	12,5
A la gente que copia le va mejor	6,4	11,4	8,5
Lo hago porque todos lo hacen	3,6	8,1	5,5
Nunca creí que era un problema	7,4	15,5	10,9
Creo que al profesor no le importa	6,9	18,7	11,9
Creo que el profesor no se da cuenta	7,3	17,5	11,6
No me queda tiempo para hacerlo de otra manera	15,9	29,9	21,9

Tabla 14: Porcentaje de alumnos que considera la justificación (para el plagio) "Buena" o "Muy Buena" vs. Frecuencia de copia

4. Conclusiones

El principal objetivo de este estudio fue identificar las percepciones y prácticas de estudiantes de educación media y superior en torno al plagio. Específicamente buscamos describir como los estudiantes entienden el plagio, cuáles son sus prácticas, bajo qué condiciones plagian y cómo justifican esta práctica. Paralelamente nuestro interés fue relacionar estas prácticas con el acceso y uso de internet y otras tecnologías de la información.

De los resultados entregados por la encuesta podemos establecer que el plagio es una práctica extendida entre los estudiantes. Cerca de un 55 % de los estudiantes de educación media declaran haber copiado y pegado información de internet sin citar la fuente, mientras que un 42 % de los estudiantes de

educación superior declaran lo mismo. Esta práctica parece estar justificada por razones de orden pragmático, que se sustentan en alguna medida en una falta de conocimiento respecto de la noción de plagio y su aplicabilidad por parte de los estudiantes. Además, el plagio entre los estudiantes se ve facilitado por una masificación del acceso a los computadores e internet, que ha llegado a ser la principal fuente de información de los alumnos, por sobre medios tradicionales de consulta bibliográfica como los libros y enciclopedias.

Confirmando la evidencia presentada por otros estudios, el factor que pareciera influir más ampliamente en la incidencia de plagio es la creencia por parte de los alumnos que sus pares han copiado, lo que da cuenta de una cierta legitimidad o naturalización de la práctica del plagio entre los estudiantes y posiblemente dentro del sistema educativo en su conjunto que normaliza el plagio como una práctica habitual entre los estudiantes. De esta manera los estudiantes que creen que sus compañeros plagiaron el año anterior son más proclives a plagiar en sus trabajos escritos. Por otro lado, entre quienes plagian, las principales razones esgrimidas son de orden pragmático tales como el ahorro de tiempo o escases de tiempo para cumplir con todos sus compromisos.

Ambos grupos de estudiantes tienen un acceso masivo al computador e internet como herramientas de trabajo, 80 % de los estudiantes de educación media y 95 % de los estudiantes de educación superior. Tanto para estudiantes de educación media como superior, internet es la principal fuente de información y parece haber reemplazado a los libros y otras fuentes de información tradicionalmente utilizadas en el marco del trabajo escolar y académico. Esta tendencia es ligeramente menos marcada para el caso de los estudiantes de educación superior, los cuáles probablemente por facilidades de acceso utilizan libros como fuente de información de manera habitual.

Al momento de buscar información en la Web, la gran mayoría de los alumnos, tanto de educación media como superior, utilizan buscadores genéricos como *Google* como su principal fuente de información, seguida de sitios como *Wikipedia*. Por su parte, los sitios de contenido educativo complementarios, como el sitio de la tradicional revista *Icarito* o *educarchile.cl*; y sitios de tareas y trabajos ya escritos, como *El Rincón del Vago*, son utilizados con menor frecuencia por los alumnos. Este patrón es el mismo para los estudiantes de educación superior pero más acentuada a favor de la utilización de buscadores generales y la no utilización de páginas especializadas.

Plagio, definiciones y conceptos

Al momento de evaluar la claridad con que los alumnos identifican prácticas de plagio, dos conclusiones merecen ser destacadas. Primero, para la mayoría de los estudiantes, no existe total claridad respecto de qué constituye

una práctica de plagio. Esta confusión es aún más marcada cuando las situaciones descritas son menos absolutas. Segundo, el nivel de confusión respecto de las definiciones da cuenta de los problemas de los alumnos para trabajar con fuentes secundarias de información. Esto es particularmente serio en un contexto en el que, tal como se ha descrito, los alumnos acceden a información por canales poco estructurados como son los buscadores genéricos, donde la cantidad de información secundaria disponible *a la mano* crece continuamente y requiere de una evaluación rigurosa sobre su veracidad y calidad.

Asociado a estos resultados, surge la pregunta por los niveles de información con los que cuentan los alumnos sobre prácticas apropiadas de cita para realizar trabajos escritos con el fin de descartar la posibilidad de que exista plagio accidental basado en la ignorancia de los estudiantes sobre el tema. Los resultados de la encuesta señalan que, si bien el porcentaje de alumnos de educación media que han recibido formación respecto de cómo citar no es despreciable (casi un 50 %), existe un desbalance entre el nivel de información acerca del carácter inadecuado de la copia y el conocimiento de prácticas de uso de información que eviten la copia o que faciliten el uso de referencias de manera adecuada. Es decir, los alumnos parecen saber cómo buscar información, y conocen del carácter negativo del plagio, pero parecen carecer de las herramientas que les permitan citar las fuentes de manera adecuada. Por otro lado, para el caso de los estudiantes de educación superior, el porcentaje de alumnos que han recibido formación respecto de cómo citar es menor a un 50 %, lo que podría ser aún insificiente para intentar disminuir el plagio.

Finalmente, los resultados de la encuesta dan cuenta de manera fehaciente que la práctica de plagio es una realidad generalizada entre los estudiantes de educación superior y media, los cuales tienen un acceso casi universal a internet y a fuentes de información digital frente a los cuáles parecieran no recibir la instrucción suficiente para mediar con la creciente cantidad y variabilidad en la calidad de la información que reciben.

Referencias

[1] Bordigon, F., Tolosa, R.A. Rodríguez y Peri, J. Primeras Experiencias en la detección de Plagio en el Ambiente Educativo. *Actas, Primera Jornada de Educación en Informática y TICS en Argentina*,pp.97-104, 2003.

[2] Braumoeller B. y Gaines, B.. Actions Do Speak Louder than Words: Deterring Plagiarism with the Use of Plagiarism-Detection Software. *PS: Political Science and Politics*, 34:835-839, 2001.

[3] Castells, M. The Rise of the Network Society: The Information Age. *Economy, Society, and Culture*. Volume I. Wiley-Blackwell, 1996.

[4] Ertmer, P. Teacher pedagogical beliefs: The final frontier in our quest for technology integration?. *Educational Technology Research and Development*, 53:25-39, 2005.

[5] Jaffee, D. Virtual Transformation: Web-Based Technology and Pedagogical Change. *Teaching Sociology* 31:227-236, 2003.

[6] Kock, N. y Davison, R. Dealing with Plagiarism in the Information Systems Research Community: A Look at Factors That Drive Plagiarism and Ways to Address Them. *MIS Quarterly* 27:511-532, 2003.

[7] McCabe, Ma. y L.K. Treviño, L. Academic Dishonesty: Honor Codes and Other Contextual Inuences. *The Journal of Higher Education* 64:522-538, 1993.

[8] McCabe, D. y L.K. Treviño What We Know about Cheating in College: Longitudinal Trends and Recent Developments. *Change* 28:28-33,1996.

[9] McCabe, D. y Treviño, L. Individual and Contextual Inuences on Academic Dishonesty: A Multicampus Investigation. *Research in Higher Education* 38:379-396, 1997.

[10] McCabe, D., Treviño, L. y Butterfield, K. Honor Codes and Other Contextual Inuences on Academic Integrity: A Replication and Extension to Modified Honor Code Settings. *Research in Higher Education* 43:357-378, 2002.

[11] McCabe, D. Cheating . Why Students Do It and How We Can Help Them Stop. *American Educator*. 38-43, 2001.

[12] Norris, C., Sullivan, T., Poirot, J. y Soloway, E. No Access, No Use, No Impact: Snapshot Surveys of Educational Technology In K-12. *Journal of College Student Development* 43:374-385, 2002.

[13] Robelen, E.W. Online Anti-Plagiarism Service Sets on Court Fight. *Education Week*. 26 (36), ISSN-0277-423, 2002.

[14] Roberts, T. Student Plagiarism in an Online World: Problems and Solutions. *Idea Group Reference*, 2007.

[15] Scanlon, P.M. y Neumann, D.R. Internet Plagiarism Among College Students. *Journal of College Student Development.* 43:374-385, 2002.

[16] Wells, J. y Lewis, L. Internet Access in U.S. Public Schools and Classrooms 1994-2005. *Highlights. NCES 2007-020.* 2006. ED Pubs. P.O. Box 1398, Jessup, MD 20794- 1398. Tel: 877-433-7827; Web site: http://www.edpubs.org y `http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED494307` (Fecha de consulta: 10 de Enero de 2010).

5. Anexos

5.1. Anexo A: Diseño Muestral

El diseño muestral del estudio consideró tres universos de estudiantes:

- Estudiantes matriculados en educación media durante el año académico 2009, en establecimientos ubicados en zonas urbanas de la RM.[2]
- Estudiantes matriculados en educación media durante el año académico 2009, en establecimientos ubicados en zonas urbanas de la V Región.[3]
- Estudiantes matriculados en educación superior durante el año académico 2008, en instituciones de educación superior ubicados en la RM.[4]

La representatividad del muestreo se restringe a los estudiantes de educación media de cada región particular y a los estudiantes de educación superior de la RM.

Para cada uno de estos universos, se escogió una muestra estratificada y por conglomerados, en múltiples etapas. En el caso de los estudiantes de educación media, la muestra consideró 1.379 estudiantes, para un universo de 368.480 alumnos de educación media en la RM. En el caso de la V Región, la muestra consideró 634 estudiantes, para un universo de 97.486 alumnos de educación media.

Para el caso de los estudiantes de educación superior en la Región Metropolitana, se escogió una muestra estratificada en dos etapas (tipo de institución; Área de estudio) y luego por conglomerados (institución). La muestra consideró 1.126 estudiantes, para un universo de 384.135 estudiantes. Los tamaños muestrales han sido calculados de modo de obtener un margen de error de ± 3 puntos porcentuales, con un nivel de confianza de 95 %.

[2] Fuente: Base de datos, estudiantes matriculados 2009. Sitio web: `www.mineduc.cl`
[3] *Ibíd.*
[4] Fuente: Estudiantes Matriculados, 2008, Consejo de Educación Superior.

El procedimiento de selección de la muestra de estudiantes de **educación media** fue el siguiente:

- Se calculó el número total de alumnos en educación media durante el año académico 2009, a partir de datos de matrícula provistos por MINEDUC.
- A partir de los tamaños poblacionales, se calculó el tamaño muestral requerido para obtener márgenes de error de de ±3 puntos porcentuales, con un 95 % de confianza.
- Usando el N° identificador de cada colegio, se suplementó la información de Matrícula (provista por MINEDUC) con la información de grupo socioeconómico de los establecimientos (Base de datos, SIMCE 2008). El proceso generó un total de 254 establecimientos sin información sobre grupo socioeconómico (sobre un total de 1.491 establecimientos), que fueron luego excluidos del cálculo de la proporción de casos a asignar a cada estrato. Cabe destacar, sin embargo, que los alumnos de dichos establecimientos sí fueron considerados para calcular el tamaño muestral.
- Se dividió la muestra por estratos, en dos etapas:

  - Primera estratificación: dependencia del establecimiento

  - Segunda estratificación: nivel socioeconómico del establecimiento, según clasificación SIMCE.
- Finalmente, se escogió aleatoriamente alumnos agrupados por conglomerado (establecimientos) dentro de cada estrato.
- Para cada establecimiento, se escogió un nuevo conglomerado (curso) de modo de generar el número de casos requerido. Se escogió aleatoriamente dos cursos (de educación media) por colegio.

Por otra parte, el procedimiento de selección de la muestra de estudiantes de **educación superior** fue el siguiente:

- Se calculó el número total de alumnos matriculados en educación superior durante el año académico 2008, a partir de datos de matrícula provistos por el Consejo de Educación Superior.
- A partir de los tamaños poblacionales, se calculó el tamaño muestral requerido para obtener márgenes de error de de ±3 puntos porcentuales, con un 95 % de confianza.

- Se dividió la muestra por estratos, en una etapa:
  - Primera estratificación: tipo de institución (CFT, IP, Universidad).
  - Segunda estratificación: área de estudio (según OECD).
- En la última etapa, se escogió aleatoriamente alumnos agrupados por institución dentro de cada estrato.

Agradecimientos

Los autores agradecen el aporte del Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16)

Segmentación automática de la Provincia de Buenos Aires para el Censo Nacional Argentino 2010

Flavia Bonomo [*]
Diego Delle Donne [**]
Guillermo Durán [***]
Javier Marenco [**]

Resumen

La planificación de un censo poblacional implica diversos desafíos logísticos, uno de los cuales es determinar qué hogares debe visitar cada censista. Este problema se conoce como el problema de segmentación de viviendas, y generalmente involucra un conjunto de restricciones sobre los recorridos que cada censista puede realizar y determinados criterios de homogeneidad y uniformidad que las soluciones deben respetar. En este trabajo presentamos un enfoque basado en programación lineal entera para la resolución del problema de segmentación en las zonas urbanas y semi-urbanas de la Provincia de Buenos Aires, que fue aplicado exitosamente durante el Censo Nacional 2010 en Argentina.

Palabras clave: Censo Poblacional, Programación lineal entera, Segmentación.

[*]Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina y IMAS–CONICET, Argentina

[**]Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina e Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina y IMAS–CONICET, Argentina

[***]Depto. de Matemática, FCEyN, Universidad de Buenos Aires, Argentina, Depto. de Ingeniería Industrial, FCFM, Universidad de Chile, Chile e IMAS–CONICET, Argentina

1. Introducción

Un *censo poblacional* es el proceso de adquisición de información demográfica de un área geográfica (habitualmente un país). El objetivo principal es relevar datos estadísticos sobre los habitantes, incluyendo información sobre sus viviendas, su nivel educativo y su realidad laboral. Típicamente, un censo se realiza a lo largo de un sólo dí a declarado no laborable, y la información es recabada por *censistas* que visitan casa por casa a todos los habitantes del área censada.

Las actividades de planificación de un censo suelen comenzarse con varios años de anticipación, dado que un censo involucra un importante movimiento logístico realizado durante unas pocas horas. Una tarea crucial dentro de la planificación de un censo es determinar qué viviendas deberá visitar cada censista, problema que en este contexto se denomina *problema de segmentación de viviendas*. Se trata de un problema de *partición* del conjunto de viviendas en subconjuntos que cumplan cierto conjunto de restricciones, y puede ser de muy difícil resolución dependiendo del tipo de restricciones y de los objetivos de la partición.

Cuando esta partición se realiza a relativamente alto nivel –por ejemplo, para determinar distritos electorales dentro de un área geográfica– el problema es conocido como *problema de re-zonificación* (*redistricting problem*, en inglés), y para este tipo de problemas existe una abundante literatura [2, 3, 6, 8, 9, 10, 11]. Existen varios paquetes de software que permiten realizar este tipo de tareas en forma automática y semi-automática (por ejemplo, en [4] se presenta una herramienta *open source* implementada para el entorno de programación estadística R). Sin embargo, las restricciones presentes en los problemas de re-zonificación en general difieren de los requerimientos de un problema típico de segmentación de viviendas, con lo cual se hace necesario recurrir a algoritmos específicos para este último caso. Además de estas restricciones, se agregan al problema de segmentación de viviendas ciertos criterios de homogeneidad y uniformidad que deben ser respetados por las soluciones halladas (los mismos se detallan en la Sección). En los enfoques de resolución manual utilizados en la actualidad, estos criterios son prácticamente imposibles de respetar, ya que las soluciones dependen fuertemente de las decisiones de cada operador.

En este trabajo describimos el problema de segmentación de viviendas para el Censo Nacional 2010 en la Provincia de Buenos Aires (Argentina), y presenta-

mos una herramienta basada en programación lineal entera para su resolución. Esta herramienta fue utilizada exitósamente durante el proceso de planificación del Censo Nacional 2010 en esta provincia. No estamos al tanto de otros trabajos en la literatura de investigación de operaciones que hayan encarado el problema de segmentación de viviendas en el contexto de un censo.
El resto de este trabajo está organizado del siguiente modo. En la Sección damos una descripción general del Censo Nacional 2010 y del problema particular de segmentación de viviendas que nos ocupa. La Sección 2.1 contiene los detalles del algoritmo propuesto y menciona las principales dificultades encontradas durante su implementación. Por último, la Sección 3.1 menciona los resultados obtenidos con esta herramienta y la Sección 3.1 cierra el trabajo con conclusiones sobre el proceso realizado.

2. Descripción del problema

El territorio de Argentina está dividido en 23 *provincias* además de la Ciudad Autónoma de Buenos Aires (a la que habitualmente se refiere como Capital Federal), que tiene un status jurídico especial. A su vez, cada provincia está dividida en *partidos* o *departamentos*. A los fines del censo, cada partido se divide en conjuntos de manzanas contiguas denominados *radios censales*. Cada radio censal contiene aproximadamente 300 hogares y, dependiendo de su densidad poblacional, entre 1 y 50 manzanas. A lo largo de este trabajo nos referiremos a los radios censales simplemente como "radios".

Figura 1: Mapa de Argentina, resaltando la Provincia de Buenos Aires.

La Provincia de Buenos Aires (ver Figura 1) es la provincia de mayor superficie (307.571 km^2) y mayor cantidad de habitantes (unos 15.3 millones de

habitantes) de Argentina. No incluye a la Ciudad Autónoma de Buenos Aires pero sí incluye al Gran Buenos Aires, un cordón de unos 9 millones de habitantes que rodea al territorio de la Ciudad de Buenos Aires y conforma junto con ella un núcleo poblacional único. Salvo el Gran Buenos Aires, la Provincia de Buenos Aires es predominantemente de carácter rural, con algunos sectores definidos dedicados al turismo (sobre la costa atlántica), la explotación minera (en el sur de la provincia) y la industria metalúrgica (en el sector noreste, sobre la costa del Rí o Paraná). En contraste, el Gran Buenos Aires tiene un perfil fundamentalmente urbano e industrial. La Provincia de Buenos Aires está dividida en 134 *partidos*, dentro de los cuales se ubican 16.691 radios. En este trabajo estamos focalizados en la segmentación de estos casi 17.000 radios censales.

Dado un radio, el problema de segmentación de viviendas consiste en particionar las viviendas del radio en conjuntos disjuntos de viviendas –llamados *segmentos*– de modo tal que cada segmento sea asignado a un censista. Cada segmento debe estar compuesto por viviendas *contiguas*. Dos viviendas se consideran contiguas si están en la misma manzana y son adyacentes, o bien pertenecen a esquinas de dos manzanas distintas y dichas esquinas están enfrentadas sobre una misma calle, es decir, un segmento no puede cruzar una esquina en diagonal. La Figura 2 muestra algunos ejemplos de segmentos factibles y no factibles de acuerdo con este criterio de contigüidad. Este criterio está determinado por el manual de procedimientos censales de la Dirección Provincial de Estadísticas de Buenos Aires.

Denominamos *segmentación* a la partición de las viviendas de un radio en un conjunto de segmentos. De acuerdo con el manual de procedimientos del Censo Nacional 2010, una segmentación factible debe cumplir las siguientes restricciones:

- Cada vivienda debe pertenecer a un único segmento y el conjunto de segmentos debe cubrir a todas las viviendas del radio.
- Si un *lado* de una manzana no tiene ninguna vivienda, ese lado también debe formar parte de algún segmento. Es importante notar que denominamos "lados" a los bordes rectos de una manzana. En los trazados urbanos tí picos de las ciudades argentinas, habitualmente una manzana es un rectángulo que contiene cuatro lados, aunque no es infrecuente encontrar manzanas con morfologí as distintas.
- Cada segmento debe contener entre 32 y 40 viviendas.
- Cada censista no debe recorrer una distancia superior a cierta cota superior L, que depende de la densidad poblacional del radio en cuestión.

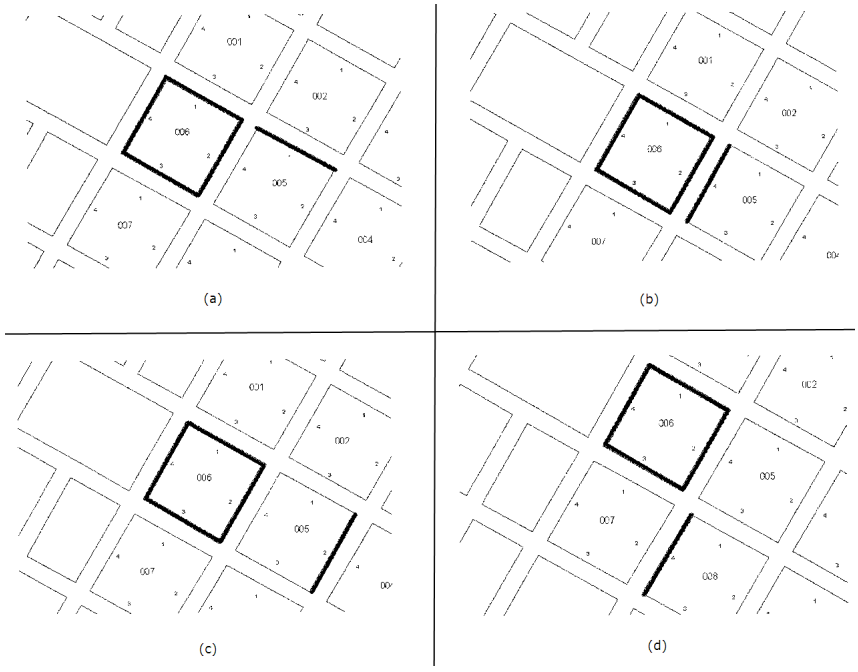


Figura 2: Los segmentos mostrados en (a) y (b) son ejemplos de segmentos factibles que cruzan la calle. El segmento mostrado en (c) no es factible dado que las viviendas no son contiguas. El segmento mostrado en (d) no es factible dado que cruza una esquina en diagonal.

- Los censistas no pueden cruzar avenidas, vías del ferrocarril ni cursos de agua.
- Un segmento debe estar contenido dentro de su radio.

La última restricción es muy importante para nuestros propósitos, dado que como un segmento no puede involucrar viviendas de dos radios distintos, entonces el problema de segmentar todas las viviendas de la provincia se reduce a 16.691 instancias individuales, una por cada radio. Por otra parte, los siguientes elementos son deseables para la segmentación, en orden de preferencia:

1. Se deben privilegiar los segmentos que consten de manzanas completas.
2. Los segmentos deben consistir de lados completos y ser tan "compactos" como sea posible. Este requerimiento de *compacidad* está definido informalmente y hace referencia a la amplitud de un segmento a lo largo de las manzanas del radio al que pertenece. Por ejemplo, un segmento constituido por una o varias manzanas completas es considerado muy com-

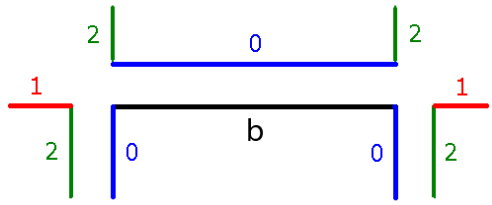


Figura 3: Niveles de adyacencia de los lados que pueden recorrerse a continuación del lado b en un segmento factible.

pacto, mientras que un segmento que recorre varias manzanas utilizando pocos lados por manzana es considerado poco compacto. El objetivo de emplear segmentos tan compactos como sea posible es el de minimizar la posibilidad de que los censistas cometan errores en sus recorridos.

3. Si no es posible contar con lados completos (por ejemplo, si un lado tiene más de 40 viviendas), se puede partir un lado y distribuirlo en más de un segmento. En este caso se debe dar preferencia a los segmentos que contengan todos los apartamentos de un mismo edificio, de modo tal que cada edificio sea atendido por un mismo censista.

4. Si el punto anterior no es posible (por ejemplo, porque un edificio tiene más de 40 apartamentos), entonces se debe dar preferencia a los segmentos que no dividan los apartamentos de un mismo piso. En otras palabras, idealmente cada piso de un edificio debe ser atendido por un mismo censista.

Como preferencia adicional, luego de recorrer un lado de una manzana es deseable o bien no cruzar la calle o bien continuar por el lado directamente enfrentado al lado recorrido. La Figura 3 especifica esta preferencia. Luego de recorrer el lado b de la figura, es preferible no cruzar la calle y continuar con los lados marcados con 0 de la misma manzana o bien cruzar al lado directamente enfrentado a b. Si esto no es posible, es preferible continuar el segmento por el lado de mismo sentido en la siguiente manzana (lados marcados con 1 en la figura). Si esto no es posible, se admite dar un giro al recorrido y continuar por alguno de los lados en las manzanas adyacentes (lados marcados con 2 en la figura).

2.1. Modelo de programación lineal entera

Sobre la base de la descripción anterior, se puede formular un modelo de programación lineal entera para este problema. Sea $\mathcal{R}$el radio a procesar, y sea $\mathbb{S}_\mathcal{R}$ el conjunto de todos los segmentos factibles sobre el radio $\mathcal{R}$. Para cada segmento $s \in \mathbb{S}_\mathcal{R}$ introducimos la variable binaria x_s, de modo tal que $x_s = 1$ si y sólo si el segmento s participa de la solución.

Con el objetivo de maximizar la compacidad de los segmentos seleccionados, proponemos la siguiente función objetivo. Dado un segmento $s \in \mathbb{S}_\mathcal{R}$, definimos su *compacidad* como $\text{comp}(s) = \frac{\text{lados}(s)}{\text{manzanas}(s)}$, donde lados($s$) y manzanas($s$) representan la cantidad de lados y manzanas, respectivamente, en el segmento s. De esta manera, un segmento que involucra una manzana rectangular completa tendrá una compacidad de 4, mientras que un segmento que involucre cuatro lados en cuatro manzanas distintas tendrá una compacidad de 1. Cabe notar que con esta definición, un segmento formado por una manzana completa logra la misma compacidad que sumar dos segmentos formados por media manzana cada uno, o bien cuatro segmentos formados por cuartos de manzana, sin embargo el primer caso es el más deseado. Para priorizar los segmentos de mayor compacidad y evitar estos casos, definimos la *valuación* de un segmento s como $\text{val}(s) = k^{\text{comp}(s)}$. Tras algunas pruebas preliminares para determinar el valor de k, se fijó este valor en 10, pues con este valor se obtuvieron los mejores resultados. Este criterio fue definido en conjunto con los responsables de la planificación del censo en la Provincia de Buenos Aires, y proporcionó resultados satisfactorios.

Sea $\mathbb{V}$ el conjunto de viviendas de $\mathcal{R}$, y sea $\mathbb{L}_0$ el conjunto de lados sin viviendas de $\mathcal{R}$. Para cada $v \in \mathbb{V}$ llamamos $S_v \subseteq \mathbb{S}_\mathcal{R}$ al conjunto de segmentos factibles que incluyen a la vivienda v, y para cada $l \in \mathbb{L}_0$ llamamos $L_l \subseteq \mathbb{S}_\mathcal{R}$ al conjunto de segmentos factibles que incluyen al lado l. Con estas definiciones, el modelo de programación lineal entera de segmentación para este problema es el siguiente:

$$\text{máx} \sum_{s \in \mathbb{S}_\mathcal{R}} \text{val(s)} \cdot x_s$$

$$\sum_{s \in S_v} x_s = 1 \qquad \forall v \in \mathbb{V} \tag{1}$$

$$\sum_{s \in L_l} x_s = 1 \qquad \forall l \in \mathbb{L}_0 \tag{2}$$

$$x_s \in \{0, 1\} \qquad \forall s \in \mathbb{S}_\mathcal{R} \tag{3}$$

Las restricciones (1) aseguran que cada vivienda es cubierta por exactamente un segmento, y las restricciones (2) aseguran que cada lado sin viviendas es recorrido por exactamente un segmento. Es importante observar que no estamos incluyendo en este modelo las preferencias en cuanto a los niveles de adyacencia utilizados en los segmentos que cruzan las calles.

Este modelo tiene en general una cantidad muy grande de variables, dado que la cantidad de segmentos factibles crece en forma exponencial a medida que aumenta la cantidad de manzanas en el radio en cuestión. Por este motivo, no resulta práctico ejecutar el modelo con todos los segmentos factibles. Una opción natural es recurrir a un algoritmo de generación de columnas (ver [5] para mayores detalles sobre esta técnica), aunque en este caso no resulta claro a priori cómo resolver satisfactoriamente el subproblema de generación de columnas (generar segmentos válidos con costos reducidos positivos, en nuestro caso) sin recurrir a heurísticas o búsquedas exhaustivas, además de que no se tendrían en cuenta las preferencias en cuanto a los niveles de adyacencia utilizados al cruzar la calle. Estas preferencias se podrían incorporar en la función objetivo a través de la función de valoración de los segmentos, aunque esto agregaría parámetros adicionales que deberían ser ajustados para obtener resultados satisfactorios.

Por estos motivos, en el presente trabajo implementamos un enfoque basado en la resolución secuencial de este modelo para conjuntos cada vez más grandes de segmentos, teniendo en cuenta explícitamente los niveles de adyacencia utilizados por los segmentos. Este esquema nos permitió hallar soluciones satisfactorias, de acuerdo con los resultados reportados en la Sección 3.1.

3. El algoritmo de segmentación

Describimos en esta sección el algoritmo que implementamos para encarar la resolución del problema de segmentación de viviendas. Decimos que un segmento se encuentra *excedido* si contiene más de 40 viviendas o su longitud supera el límite L. Decimos que un segmento es *factible* si no está excedido y tiene al menos 32 viviendas. En el contexto del algoritmo, los segmentos no son necesariamente factibles. Para $\delta \in \{0, 1, 2\}$, decimos que un segmento es *δ-conexo* si sus lados están conectados por adyacencias de tipo a lo sumo δ, de acuerdo con la especificación de la Figura 3.

El Algoritmo 1 ilustra el procedimiento propuesto, tomando los datos geográficos del radio como datos de entrada y el valor $\delta \in \{0, 1, 2\}$ como parámetro. Para $i \geq 1$, denominamos S_i al conjunto de segmentos no excedidos (aunque no necesariamente factibles) que involucran no más de i manza-

nas. En la i-ésima iteración, el Algoritmo 1 ejecuta el modelo de programación lineal entera de segmentación utilizando solamente los segmentos factibles S'_i del conjunto S_i. Si el modelo no tiene solución factible, se continúa iterando hasta que el modelo tenga solución factible, o S_i no varíe con respecto a S_{i-1}, o bien hasta llegar a un límite MI de iteraciones especificado de antemano. En cuanto el modelo es factible, el algoritmo termina retornando la solución obtenida por el modelo.

Algoritmo 1 Algoritmo de segmentación para nivel δ de adyacencias.

```
 1: S_b ← {}                                   // conjunto base
 2: para toda manzana q hacer
 3:     S_b ← S_b ∪ {segmentos de q no excedidos}
 4: fin (para)
 5: i ← 1
 6: S_i ← S_b
 7: repetir
 8:     Ejecutar el modelo de PLE de segmentación con los segmentos factibles de S_i
 9:     si hay solución entonces
10:         Terminar (con solución)
11:     fin (si)
12:     S_{i+1} ← S_i
13:     para todo (s_i, s_b) ∈ S_i × S_b hacer
14:         si s_i ∪ s_b es un segmento δ-conexo no excedido entonces
15:             S_{i+1} ← S_{i+1} ∪ {(s_i ∪ s_b)}
16:         fin (si)
17:     fin (para)
18:     si S_{i+1} = S_i entonces
19:         Terminar (sin solución)
20:     fin (si)
21:     i ← i + 1
22: hasta que i > MI
23: Terminar (sin solución)
```

Es importante observar que la solución obtenida por este algoritmo puede no ser óptima para el modelo de PLE de segmentación utilizando todos los segmentos, dado que se ejecuta este modelo con un subconjunto de segmentos $S'_i \subseteq \mathbb{S}_{\mathcal{R}}$. Sin embargo, este procedimiento secuencial permite tener en cuenta de una manera muy natural las preferencias sobre la compacidad de los segmentos seleccionados, aumentando la posibilidad de encontrar primero las soluciones preferenciales. Vale aclarar que la solución obtenida es factible para dicho modelo.

En las líneas 2–4, el algoritmo genera el conjunto base de segmentos S_b utilizando lados de manzanas completos. En este paso, se generan para cada manzana todos los posibles segmentos no excedidos contenidos en la manzana.

Con el objetivo de hallar soluciones formadas por segmentos lo más compactos posible, se ejecuta el Algoritmo 1 secuencialmente para $\delta = 0, 1, 2$, interrumpiendo el procedimiento cuando se encuentra la primera solución factible.

Si luego de todo el proceso mencionado no se encuentra solución factible, se activa la opción de partir lados. Para ello, se introduce un parámetro P que indica el número máximo de viviendas que puede tener un lado y se fragmentan en dos o más partes los lados que superen en viviendas a este valor. Cada una de estas *partes* debe tener a lo sumo P viviendas. El algoritmo de fragmentación de lados es un procedimiento goloso, e intenta dejar en una misma parte los apartamentos de un mismo edificio. Para ello, comenzando desde una de las esquinas del lado a partir, se construye una *parte* del lado recorriendo el mismo hasta encontrar un edificio o bien hasta juntar P viviendas. En cualquiera de los dos casos, se termina la *parte* actual y se comienza una nueva. Siguiendo el mismo procedimiento se obtiene el conjunto de partes correspondientes al lado partido. Finalmente, para estos lados partidos se especifican nuevas adyacencias, de acuerdo con la Figura 4. Vale aclarar que

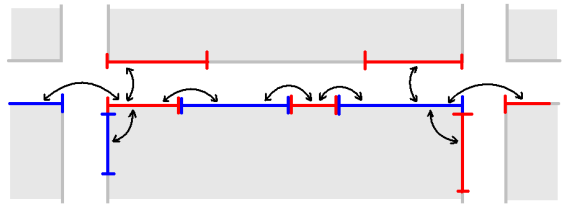


Figura 4: Adyacencias para lados partidos.

mientras más chico sea P, mayor será la cantidad de lados (i.e., partes) de las manzanas del radio y mayor será el tamaño del conjunto base S_b.

Una vez partidos los lados de acuerdo con la descripción anterior, se ejecuta el Algoritmo 1 nuevamente para $\delta = 0, 1, 2$, intentando encontrar una solución factible. Si aún así no se encuentra una solución factible, se reduce el valor de P y se repite el proceso. El Algoritmo 2 ilustra este procedimiento recibiendo como parámetro una lista PL que contiene los sucesivos valores a utilizar para el parámetro P. Vale aclarar que estos valores se utilizan en orden decreciente, ya que mientras mayor sea el valor de P, menor fragmentación tendrán las soluciones obtenidas. Si luego de todo este procedimiento no se encuentra una solución factible, el procedimiento termina informando al usuario que no se pudo encontrar una segmentación. Cabe mencionar que inicialmente el valor de P es menor o igual a 40 (veremos más adelante que estos valores dependen de la densidad poblacional del radio a resolver).

Algoritmo 2 Algoritmo de segmentación

1: **para** cada $P \in PL$ **hacer**
2: Partir los lados que superen en viviendas el valor P.
3: **para** $\delta = 0$ **hasta** 2 **hacer**
4: Ejecutar el Algoritmo 1 para δ
5: **si** el Algoritmo 1 encontró una solución **entonces**
6: Retornar la solución y terminar
7: **fin (si)**
8: **fin (para)**
9: **fin (para)**
10: Terminar (sin solución)

3.1. Mejoras para radios con baja densidad poblacional

En las etapas de prueba y ajustes preliminares, el Algoritmo 2 permitió resolver una gran cantidad de radios urbanos de la Provincia de Buenos Aires. Sin embargo, este algoritmo tuvo serios problemas para resolver radios semi-urbanos con densidades poblacionales relativamente bajas. Por ejemplo, la Figura 5 muestra un radio de la Ciudad de Olavarría para el cual este algoritmo no logró encontrar una solución factible. El principal problema en este tipo de radios es que tienen un gran número de manzanas sin viviendas o muy poco pobladas, y esto hace que el algoritmo requiera alcanzar un alto número de iteraciones para obtener suficientes segmentos factibles como para poder cubrir el radio entero. Por otro lado, al tener muy pocas viviendas por manzana, la cantidad de segmentos no excedidos es excesivamente grande. Por ejemplo, el conjunto S'_7 para el radio de la Figura 5 está formado por más de 100,000 segmentos, lo que hace que tanto la generación de los segmentos como la resolución del modelo tome demasiado tiempo. La combinación de estas dos características hace que sea imposible llevar a la práctica el algoritmo implementado.

Sin embargo, el hecho de que un radio esté muy poco poblado hace que no sea necesario manejar tantos segmentos base, ya que las restricciones para combinarlos son menos prohibitivas (teniendo en cuenta los límites inferior y superior de viviendas por segmento). Así, para solucionar los problemas mencionados, agregamos al algoritmo las siguientes mejoras:

- En primer lugar, las manzanas que tienen muy pocas viviendas son consideradas como un único segmento no excedido al momento de agregar segmentos al conjunto base S_b (línea 3 del Algoritmo 1). Para ello se agrega un parámetro MP que indica el mínimo número de viviendas que debe tener una manzana para generar varios segmentos base a par-

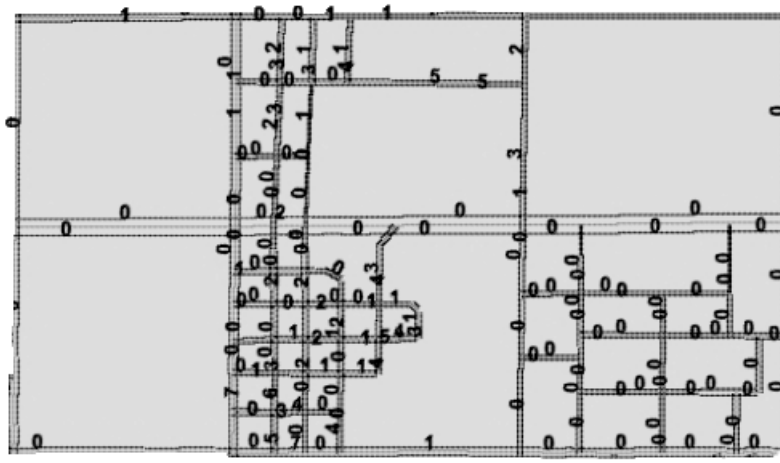


Figura 5: Radio semi-urbano de la Ciudad de Olavarría para el cual el Algoritmo 2 tiene una performance inaceptable. Los números en la figura indican la cantidad de viviendas en cada lado.

tir de ella. Es decir, si una manzana no alcanza este valor de viviendas, entonces sólo se agrega a S_b el segmento correspondiente a la manzana completa, en lugar de considerar todos los segmentos posibles de la manzana. Esto contribuye a reducir considerablemente el tamaño de S_b en radios de baja densidad de población.

- Una vez generados los segmentos base S_b, si alguno de ellos tiene pocas viviendas, se lo combina arbitrariamente con algún segmento adyacente formando un único segmento base con los dos. Para ello se agrega un parámetro MH que indica el mínimo número de viviendas que deben tener los segmentos de S_b. Este procesamiento del conjunto S_b se agrega al Algoritmo 1 a continuación de la línea 4.

Estos parámetros deben ser manejados con cuidado ya que según el radio a resolver, valores erróneos pueden incrementar los tiempos de ejecución (por ejemplo, valores muy bajos para P) o bien comprometer la factibilidad del modelo (por ejemplo, valores muy altos para MP y MH).

4. Resultados computacionales

Dada la gran cantidad de parámetros definidos y su relevancia para los tiempos de ejecución del algoritmo, fue necesario realizar una experimentación preliminar con diversos radios de prueba, para definir el mejor juego de parámetros

para cada tipo de radio. Como resultado de esta experimentación se resolvió clasificar los radios según su densidad poblacional en tres categorías: *urbanos* (hasta 10 manzanas), *semi-urbanos* (entre 11 y 30 manzanas) y *rurales* (más de 30 manzanas). La Tabla 1 muestra los valores seleccionados para los parámetros, de acuerdo con las características de cada radio.

	Parám.	Urbanos	Semi-urbanos	Rurales
Cantidad máxima de iteraciones en la generación de segmentos	MI	4	7	9
Número mínimo de viviendas para partir una manzana	MP	1	2	10
Número mínimo de viviendas en los segmentos base (si no, se agrupan con otros)	MH	0	1	5
Número máximo de viviendas por parte	PL	[32, 16, 10]	[32, 16]	[40, 32, 20]
Límite de tiempo para la ejecución del modelo de IP (seg)	MT	60	60	120

Tabla 1: Valores utilizados para los parámetros del algoritmo según el tipo de radio.

Los algoritmos mencionados en la sección anterior fueron codificados en C++ y los modelos de programación lineal entera fueron resueltos con Cplex 12.1 [7]. Los datos de manzanas, lados y viviendas se tomaron de la base de datos geográfica de la Provincia de Buenos Aires, implementando en un sistema de información geográfica las interfaces necesarias para exportar la información e importar y visualizar las segmentaciones obtenidas por nuestro algoritmo.

La generación de los segmentos requiere de un tiempo relativamente corto, siendo el peor caso de unos 2 minutos con los parámetros especificados en la Tabla 1. Más del 99 % de los modelos de programación lineal entera se resuelve en pocos segundos, dado que la relajación lineal resulta muy ajustada y la primera solución factible hallada por Cplex suele ser óptima. En pocos casos se llega al límite de tiempo MT especificado en la Tabla 1 con solución sub-óptima, la cual se toma dentro del Algoritmo 1 como la segmentación del radio.

El proceso de segmentación en la Provincia de Buenos Aires para el Censo Nacional 2001 (el anterior censo realizado en Argentina) se realizó manualmente, demandando 25 operadores a tiempo completo que trabajaron durante 30 días (es decir, unas 6,000 horas-hombre). En cambio, el Censo Nacional 2010 fue la primera oportunidad en la que la segmentación de viviendas en la Provincia de Buenos Aires se realizó por medio de herramientas computacionales

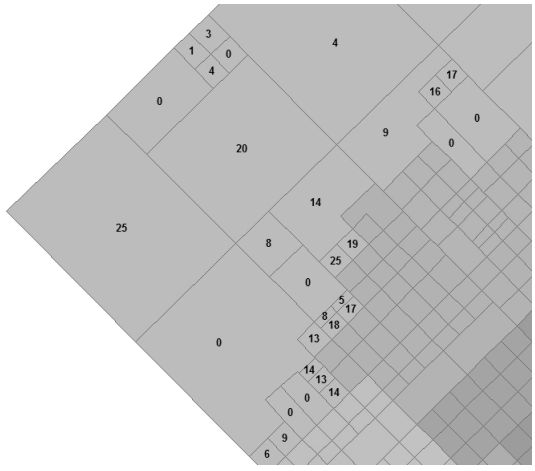


Figura 6: Radio semi-urbano para el cual el Algoritmo 2 no puede encontrar solución factible.

automáticas. Nuestro algoritmo obtuvo segmentaciones satisfactorias para el 96 % de los radios en aproximadamente 320 horas de CPU sobre un computador con procesador Intel Celeron© a 2.4 GHz y memoria RAM de 2 GB (el equivalente a un tiempo menor a 1 día en un cluster de 15 computadores).

En aproximadamente unos 600 radios (el 4 % del total) el algoritmo no encontró una solución factible en forma automática. Estos radios fueron resueltos utilizando la herramienta pero relajando levemente las restricciones del problema, o bien en forma manual. Por ejemplo, la Figura 6 corresponde a un radio semi-urbano muy poco poblado, en el cual las manzanas con pocas viviendas se agrupan entre sí para formar bloques de manzanas indivisibles (según la agrupación de manzanas descripta en la Sub-Sección 3.1). Este radio contiene tres manzanas contiguas con 14, 13 y 14 viviendas, respectivamente (ubicadas en la mitad inferior de la figura), las cuales están circundadas por manzanas escasamente pobladas, y por lo tanto probablemente agrupadas en un único bloque, "aislando" a estas tres manzanas. Estas manzanas suman 41 viviendas, con lo cual no es posible ubicarlas en un único segmento (superan el límite de 40 viviendas para un segmento) ni tampoco armar dos segmentos con ellas (segmentos con menos de 32 viviendas). Este radio particular se resolvió en pocos segundos relajando en una unidad el límite superior de 40 viviendas por segmento.

Muchos de los radios no resueltos por nuestro algoritmo tienen características similares al radio de la Figura 6. Es decir, se trata de radios predominantemente poco poblados pero con zonas densas en las fronteras (usualmente en las fronteras linderas con radios de mayor densidad). Al comentar estas carac-

terísticas con el equipo responsable de la planificación del censo, la respuesta obtenida fue que estos radios están incorrectamente delimitados, ya que no es deseable que un radio contenga zonas con densidades poblacionales muy distintas. Estos efectos son causados por el crecimiento poblacional.

5. Conclusiones

El algoritmo descrito en este trabajo permitió realizar el proceso de segmentación de viviendas de la Provincia de Buenos Aires dentro de los plazos que imponía la planificación del censo. Cabe destacar que el tiempo de desarrollo impuesto para la resolución del problema fue de tan sólo dos meses. A diferencia del proceso de segmentación manual empleado en el censo anterior, en el que cada operador podía introducir un sesgo en la segmentación realizada, el proceso automático a cargo del algoritmo permitió generar una segmentación con criterios uniformes para toda la provincia, contribuyendo así a una división del trabajo pareja entre los censistas. Además, los tiempos de procesamiento disminuyeron considerablemente con relación a la operatoria manual.

La performance del algoritmo demostró ser muy sensible a la parametrización utilizada. Con un conjunto adecuado de parámetros la resolución puede demorar unos pocos segundos, mientras que un juego de parámetros mal seleccionado puede comprometer la factibilidad del problema o bien llevar la cantidad de segmentos a varias centenas de miles, haciendo que la resolución demore varias horas. La clasificación de los radios de acuerdo con su densidad nos permitió manejar adecuadamente estos parámetros. Por último, el procedimiento secuencial dado por los algoritmos descritos fue fundamental para cumplir con las preferencias en cuando a las características deseables para los segmentos.

No obstante los resultados obtenidos, los algoritmos implementados en este trabajo ofrecen una solución heurística al problema. Como trabajo futuro, sería interesante profundizar en el enfoque de generación de columnas para la resolución exacta del modelo de programación entera planteado. El desafío mayor en este enfoque es la resolución del subproblema de generación de columnas.

Considerando que los radios censales tienen aproximadamente 300 viviendas y que cada censista recorre entre 32 y 40 viviendas, la cantidad de censistas necesaria para cada radio varía muy poco (i.e., entre 8 y 10 censistas aproxima-

damente). Por lo tanto, en cada instancia no parece tener sentido un objetivo que involucre la minimización de los censistas requeridos. Sin embargo, considerando que en la provincia hay casi 17.000 radios censales, éste puede ser un punto interesante a optimizar. Este aspecto del problema no fue abordado en este trabajo, pero puede ser un trabajo interesante a futuro.

En lo que hace a la opinión de los usuarios de la herramienta desarrollada, el responsable de los sistemas de información geográfica para el Censo Nacional 2010 en la Provincia de Buenos Aires, Fernando Aliaga, comenta que "el uso de esta herramienta computacional nos permitió una segmentación homogénea con criterios de compacidad uniformes, a diferencia de la segmentación manual que depende en gran medida de las decisiones de los operadores"[1]. El Censo Nacional 2010 se realizó exitosamente el 27 de Octubre de 2010, y fue calificado por las autoridades de la Provincia de Buenos Aires como un éxito organizativo [12].

Agradecimientos: Los autores agradecen al responsable de los sistemas de información geográfica para el Censo Nacional 2010 en la Provincia de Buenos Aires, Fernando Aliaga, y a su equipo de trabajo por su colaboración en varios aspectos fundamentales para el desarrollo de este trabajo. Este trabajo fue parcialmente financiado por ANPCyT PICT-2007-00518 y PICT-2007-00533, CONICET PIP 112-200901-00178, y UBACyT 20020090300094 y 20020100100980 (Argentina), FONDECyT 110797 e Instituto de Ciencias Milenio "Sistemas Complejos de Ingeniería" (Chile).

Referencias

[1] Aliaga, F., Comunicación personal, Noviembre. 2010.

[2] Altman, M., Is Automation the Answer: The Computational Complexity of Automated Redistricting, Rutgers Computer and Law Technology Journal 23(1) 81–141. 1997.

[3] Altman, M., MacDonald, K., and McDonald, M.P., From Crayons to Computers: The Evolution of Computer Use in Redistricting, Social Science Computer Review 23(3) 334–346. 2005.

[4] Altman, M. and McDonald, M.P., Bard: Better Automated Redistricting, Journal of Statistical Software 31(3). 2009.

[5] Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. P. and Vance, P. H., Branch-and-Price: Column Generation for Solving Huge Integer Programs, Operations Research 46 316–329. 1998.

[6] Bozkaya, B., Erkut, E., and Laporte, G., A tabu search heuristic and adaptive memory procedure for political districting, European Journal of Operational Research 144(1) 12–26. 2003.

[7] IBM ILOG, User's Manual for CPLEX. 2009.

[8] Flesichmann, B. and Paraschis, J.N., Solving a large scale districting problem: a case report, Comput. Oper. Res. 15(6) 521–533. 1998.

[9] Garfinkel, R.S. and Nemhauser, G.L., Optimal Political Districting by Implicit Enumeration Techniques, Management Science 16(8) B495–B508. 1970.

[10] Helbig, R.E., Orr, P.K., and Roediger, R.R., Political redistricting by computer, Commun. ACM 15(8) 735–741. 1972.

[11] Hess, S.W., Weaver, J.B., Siegfeldt, H.J., Whelan, J. N., and Zitlau, P.A., Nonpartisan Political Redistricting by Computer, Operations Research 13(6) 998–1006. 1965.

[12] La voz de Tandil, Se censó más del 95 % de las viviendas en la provincia. Recuperado el 15 de Noviembre de 2010. `http://www.lavozdetandil.com.ar/ampliar_nota.php?id_n=20090`. 2010.

Programación de Grúas para Mantención y Construcción de Buques en un Astillero Naval. Uso de Modelo Matemático.

Marcelo Guiñez [*]
Lorena Pradenas R [**]
Eliseo Melgarejo [**]

Resumen

En el presente estudio se desarrolla una propuesta de solución para el problema de asignación de grúas a proyectos o buques para la: reparación, mantención, modernización, apoyo y construcción, en un astillero naval. Se establecen los antecedentes generales del tema, y las justificaciones de éste, determinando la necesidad de proponer un modelo matemático diseñado para el problema particular, detallando y explicando todas las variables, parámetros y restricciones que presenta la solución, teniendo una alternativa al actual sistema que depende de la experiencia del programador. Además se proporciona la solución implementada en el programa ILOG CPLEX utilizando Microsoft Office Excel como plataforma de datos, obteniendo resultados para un caso particular de instancias.

Palabras Clave: Programación de Grúas, Astillero, PLEM, Aplicación a Problemas Reales.

[*]DII- Universidad de Concepción, Chile y ASMAR-Talcahuano-Chile
[**]DII- Universidad de Concepción, Chile

1. Introducción

Un astillero naval es una empresa que fabrica y repara buques comerciales y navales. Dentro de las operaciones de este tipo de empresa, la programación y distribución de grúas es una de las actividades relevantes. Cada requerimiento de grúas se realiza a través de una plataforma informática solicitando la información de entrada con 24 horas de anticipación. El jefe de turno realiza la programación de forma manual. Además de los pedidos de producción, existen requerimientos de especial prioridad los cuales deben ser también satisfechos. Algunos de los inconvenientes que se observan en la programación de grúas son:

- Programación basada en la experiencia del Jefe Turno, que no siempre es la más adecuada.
- Programación por prioridad del solicitante con mayor insistencia o mayor jerarquía.
- Asignación de grúas sin considerar el mayor beneficio de ingresos para el Astillero.
- Política de subcontratación de grúas faltantes tiene deficiencias en el orden de generar mayor ingresos al Astillero.

En el astillero se disponen de varios tipos de grúas y entre otras son: Grúas portales, grúas flotantes y grúas rodantes. En particular la grúa flotante corresponde a una embarcación con dotación propia. En caso necesario se debe subcontratar a terceros grúas lo que aumenta los costos de operaciones. Prácticamente todos los procesos de la empresa requieren obligatoriamente el uso de grúas. Algunos de estos procesos son:

- Transporte, en tareas de reparación, instalación, ubicación y movimientos de equipos de apoyo a la producción.
- Cambio de equipos de las naves (motores, compresores, ventiladores, turbinas, etc.)
- Mantención e inspección de sistemas y equipos.
- Procesos de limpieza y en órdenes de: Retiro de escombros, arena, granalla, maderas, recubrimientos, estructuras, desechos industriales, etc.

Por todo lo anterior, las grúas son un apoyo logístico fundamental para el proceso productivo en un astillero. En una programación ineficiente de éstas, se traduce en:

- Atraso de proyectos.
- Pérdida de recursos por paralización de obras, por ejemplo para eventualidades que signifiquen bloqueos de algunos intervalos de tiempo.
- Dilatar los procesos productivos.
- Reclamos e insatisfacción de los clientes.

A continuación se realiza un breve análisis de la literatura especializada y que proporciona algunas herramientas matemáticas y algorítmicas para resolver problemas de este tipo. En particular el problema de programación de grúas en terminales portuarios ha sido un tema tratado por los investigadores, los cuales se empeñan en hacer un mejor uso de las grúas para así mejorar el funcionamiento del puerto acelerando el flujo de naves que atracan en él, siendo ésta una de las medidas de desempeño más importantes para medir la calidad del Terminal portuario.

Al igual que en los puertos, los astilleros también disponen de un número limitado de grúas las cuales deben realizar una serie de trabajos contra el tiempo. En [1], se examina la programación de la grúa para los puertos, se inicia con un caso determinístico simple y lo utilizan como un elemento básico para desarrollar una comprensión del problema dinámico. En [5], se estudia la programación de grúas en puertos con restricciones espaciales y de separación. La grúas no se pueden cruzar y existe una distancia mínima entre ellas y los trabajos no pueden ser realizados simultáneamente. El objetivo es encontrar un apareamiento máximo entre grúa y trabajo, usan programación dinámica, una heurística tabú y una búsqueda local. El trabajo de [3], se programan grúas de muelles, que son obviamente los equipos más importantes en un terminal portuario, proponen un modelo de Programación Lineal Entera Mixta (MIPL, del inglés *Mixed Integer Linear Programming*), considerando restricciones relativas a la operación y proponiendo una solución basada en branch and bound y una heurística GRASP para situaciones donde *branch and bound* no puede operar. [7], propone programar múltiples grúas de patios que realizan diversos trabajos con diferentes tiempos de término en una zona con un solo carril de viaje bidimensional, se trabajó con un modelo MILP y una heurística dinámica entregando buenos resultados para cotas inferiores.

En el estudio de [4], se desarrollan modelos para programar grúas de yarda (enrieladas). En [9] muestran una heurística de búsqueda tabú para el problema de programación cuadrática y grúas en un muelle, el problema consiste

en un número fijo de grúas de muelle para carga y descarga de contenedores hacia y desde un buque. En [2], se propone un modelo que busca disminuir la tardanza de los trabajos, así como también disminuir el número de trabajos atrasados y corresponde a la fuente principal del presente estudio. Para esto usan MIPL, acompañado de la programación por restricciones. Esta propuesta híbrida resulta ser mucho más rápida en su resolución que los modelos que adoptan solo una de las herramientas por sí misma.
A su vez en [10], también presentan un problema de programación de trabajos que deben ser ejecutados y posteriormente un problema de asignación de grúas a estos trabajos también en un astillero, utilizando instancias generadas arbitrariamente.
Finalmente [6], analiza el problema de la programación grúas de muelle que se utilizan en las terminales portuarias de contenedores por mar para cargar y descargar contenedores. Se presenta un modelo de y también se propone un procedimiento heurístico de solución.
En general la mayoría de los trabajos relacionados con el tema, están formulados como Job Shop considerando, un ambiente donde hay trabajos que con diferentes operaciones y siguen distintas rutas por las máquinas o centros de trabajo. Al contrario del problema que enfrentamos nosotros, el cual se asemeja más a un ambiente de máquinas paralelas. Aquí una tarea puede ser ejecutada por cualquiera de las máquinas (grúas) que se encuentran en el astillero, con un tiempo de proceso que no será modificado por elegir una grúa u otra. El caso en estudio se desarrolla bajo este ambiente.
Basado en el problema a tratar y la revisión bibliográfica realizada en éste estudio se plantea un diseño e implementación computacional de un MIPL que permita apoyar y mejorar la programación de grúas en un astillero naval.

2. Modelo matemático propuesto

Se propone un modelo matemático que busca minimizar la tardanza sujeta a la disponibilidad horaria de las grúas y al tiempo de utilización de las mismas. Por otro lado, se presenta un criterio de priorización en donde las solicitudes son consideradas según un criterio predefinido, para ello existe un factor de prioridad (ver Tabla 1), considerando así un modelo matemático con una función objetivo que consta de minimización de las tardanzas ponderada.

Sean los siguientes conjuntos:

$J = 1, \ldots, n$, conjunto de solicitudes o trabajos o tareas a programar.

$I = 1, \ldots, m$, conjunto de grúas disponibles.

$I_s = 1, \ldots, s$, con $I_s \subseteq I$; conjunto de grúas subcontratadas disponibles.

$T' = 1, \ldots, T$, conjunto de periodos disponibles, donde $t = 1$ es el instante 00:00 horas, $t = 2$ es el instante 01:00 horas, $t = 24$ es el instante 23:00 horas.

$B' = 1, \ldots, B$, conjunto de periodos bloqueados.

Sean las siguientes variables de decisión:

L_j : Indica si la tarea $j \in J$, está o no atrasada, tiene valor 1, si a tarea $j \in J$ está atrasada o valor 0 en caso contrario.

x_{ijt} : Identifica si se asigna la grúa $i \in I$ a la tarea o solicitud $j \in J$, en el periodo $t \in T'$, tiene valor 1, si es la grúa $i \in I$ es asignada a la solicitud $j \in J$, en el periodo $t \in T'$ o valor 0, en caso contrario.

Considerando los siguientes parámetros:

p_j : Tiempo de proceso de solicitud $j \in J$, en horas.

d_j : Hora máxima de entrega de solicitud $j \in J$.

O_i : Capacidad en toneladas de la grúa $i \in I$.

R_j : Requerimiento de capacidad en toneladas de la solicitud $j \in J$.

D_{ij} : Tiene valor 1, si la grúa $i \in I$ puede atender la solicitud $i \in J$ o valor 0, en otro caso.

b_k : Periodo $k \in B'$ bloqueado.

W_j : Prioridad o peso de la solicitud $j \in J$. Donde pesos menores indican mayor importancia de la tarea. Los valores de

W_j van desde 1 a 10 según la Tabla 1:

Tipo de solicitud	Peso
Urgente	1
Solicitud estratégica	2
DS N°2	3
DS N°1	4
Proyectos prioritarios	5
Proyectos construcción 1	6
Proyectos construcción 2	7
Proyectos pesqueros 1	8
Proyectos pesqueros 2	9
Otros proyectos	10

Tabla 1: Prioridades de las Solicitudes

A continuación se presenta el modelo matemático propuesto:

$$\text{mín} \sum_{j \in J} L_i W_j \tag{1}$$

$$\sum_{i \in I} (t + p_j) x_{ijt} - d_j \geq TL_j; \qquad \forall j \in J, e \in T' \tag{2}$$

$$\sum_{i \in I} \sum t \in T' x_{ijt} = 1 \qquad \forall j \in J \tag{3}$$

$$\sum_{j \in J} R_j x_{ijt} \leq O_i; \qquad \forall i \in I, t \in T' \tag{4}$$

$$x_{ijt} \leq D_{ij}; \forall i \in I, j \in J, e \in T' \tag{5}$$

$$\sum_{j \in J} \sum t \in S x_{ijt} \leq 1; \qquad \forall j \in J, t \in T' \tag{6}$$

$$\text{con } S = \{t'/ma(t - p_j, 0) < t' < t\}$$

$$\sum_{j \in J} \sum t \in T' x_{ijt} p_j \geq 8; \qquad \forall i \in I_s \tag{7}$$

$$x_{ijt} = 0 \qquad \forall i \in I, j \in J, t \in T' \mid t > T - p_j \tag{8}$$

$$x_{ijt} = 0 \qquad \forall i \in I, j \in J, t \in T' \mid t > b_k - p_j, \forall k \in B' \tag{9}$$

2.1. Descripción de la Formulación Matemática

La expresión (1), corresponde a la función objetivo, la cual pretende minimizar el número total de tareas atrasadas, ponderadas de acuerdo a su clasificación de importancia.

La expresión (2), establece que una tarea es considerada atrasada si no cumple con su hora tope de entrega.

La expresión (3), establece que todas las tareas deben ser asignadas a los periodos disponibles, aun cuando estas estén atrasadas.

La expresión (4), establece que la capacidad en tonelajes de la solicitud no debe sobrepasar la capacidad de las grúas.

La expresión (5), hace referencia a la posibilidad de asignar una grúa a un trabajo sólo si ésta asignación es considerada como factible por el jefe de turno.

La expresión (6), establece que una grúa no puede ser ocupada por otra solicitud mientras está en proceso (restricción de continuidad).

La expresión (7), establece que cada tarea subcontratada debe tener un mínimo de 8 horas de uso.

La expresión (8), limita las asignaciones fuera de la jornada.

La expresión (9), corresponde a un conjunto de restricciones complementarias, cuando se desea restringir ciertos intervalos de tiempo, como por ejemplo descansos y/o periodos de colación.

3. Resultados

Se diseñaron instancias del problema las cuales pretenden mostrar las diferentes cualidades que presenta el modelo, considerando que la instancia número 2 corresponde a un caso de similar tamaño a lo cotidiano en la empresa, mientras que las demás instancias fueron elegidas arbitrariamente considerando diferentes eventualidades. La implementación se realiza en el software IBM ILOG CPLEX 12.1 en una CPU AMD Athlon II Dual Core M320 (2.1 GHz) con 1.75 GB en RAM, utilizando las condiciones *default* del algoritmo.

Descripción instancias utilizadas:

Instancias N°1: 3 grúas, 5 tareas y 24 períodos.

En la Tablas 2.a y 2.b se presentan los datos para la instancia N°1, se tienen 3 grúas y una es subcontratada. Debido a restricciones físicas la solicitud número 3 (tarea), sólo puede ser ejecutada por la grúa subcontratada,

Grúas	1	2	3
Oi	20	60	40

(a) Capacidades de las grúas para la instancia 1

Tareas o solicitudes	1	2	3	4	5
Rj	10	30	40	50	60
dj	5	5	10	8	8
Pj	2	5	5	6	6
Wj	4	4	2	9	4

(b) Características de las solicitudes o tareas para la instancia 1

Tabla 2: Instancias N°1

mientras que el resto de las tareas pueden ser ejecutadas por cualquier grúa.

Instancia N°2: 23 grúas, 50 tareas y 24 períodos.

Grúas	15, 15, 13, 28 , 50, 50, 4, 4
Oi	6, 6, 6, 6, 20, 20, 30, 30 30, 30, 40, 13, 50, 90, 200

(a) Capacidades de las Grúas para la Instancia 2

Tareas o solicitudes	
Rj	20, 25, 30,10, 5, 5, 5, 2, 40, 30, 90, 150, 2, 3, 30, 10, 15, 25, 8, 6, 6, 4, 4, 2, 2, 50, 50, 40, 5, 5, 1, 1, 1, 80, 190, 4, 4, 20, 30, 25, 3, 3, 30, 3
d_j	5, 24, 4, 6, 7, 9, 22, 12, 14, 14, 13, 15, 16, 20, 13, 20, 8, 15, 8, 21, 11, 17, 15, 16, 17, 10, 4, 18, 6, 4, 24, 19, 12, 9, 15, 24, 14, 12, 18, 10, 23, 24, 10, 6, 13, 15, 9 ,22, 6, 22
P_j	5, 15, 3, 3, 4, 2, 8, 8, 6, 11, 7, 9, 6, 12, 13, 10, 8, 4, 7, 18, 10, 12, 12, 14, 11, 9, 3, 9, 3, 2, 12, 12, 5, 6, 11, 15, 9, 4, 8, 8, 8, 5, 2, 4, 6, 6, 4, 9, 5, 7
W_j	9, 7, 8, 10, 10, 7, 1, 2, 7, 5, 6, 7, 7, 1, 1, 9, 2, 1, 2, 10, 1, 2, 3, 2, 7, 4, 1, 4, 8, 8, 2, 1, 2, 9, 4, 10, 10, 9, 4, 5, 5, 2, 4, 8, 7, 8, 7, 3, 1, 6

(b) Características de las solicitudes o tareas para la instancia 2

Tabla 3: Instancia N°2

En las Tablas 3.a y 3.b se presentan los datos para la instancia N°2. Las grúas número 21, 22 y 23 son subcontratadas y en este caso cualquier grúa puede atender cualquier solicitud.

Instancia N 3: 23 grúas, 50 tareas y 35 períodos.

Se utilizan los mismos datos de entrada que la instancia 2. Adicionalmente, basado en un caso real, debido a una ceremonia de inauguración, las actividades se detienen en el intervalo entre 13:00 - 14:00 (conjunto de restricciones número 9). Se aumentó el número de periodos a asignar a 35 (valor arbitrario) con el objetivo de generar una solución factible (cada proyecto debe comenzar, al menos dentro del periodo).
Los resultados del modelo aplicado a las instancias de prueba se observan en la Tabla 4 y en todos los casos se obtiene la solución óptima.

Instancia	Función objetivo	Trabajos Atrasados	Tiempo CPU (s)
1	4	1	0,06
2	5	2	3.10
3	35	7	3.57

Tabla 4: Resumen de resultados para instancias de prueba

4. Conclusiones

Las instancias revisadas comprueban en la práctica la eficacia en tiempo de resolución del modelo presentado, al menos para el tamaño actual de las faenas, lo que facilitaría una gran tarea (manual), que toma horas en realizar, en una tarea de segundos, sin contar la notable diferencia en la calidad alcanzada. El modelo en la práctica no se ha llevado a cabo pero si se considera realizar gestión es para poder materializarlo.

Finalmente se considera el énfasis en cumplir con la cabalidad de las restricciones prácticas impuestas por el astillero naval, sin embargo se esperaría a futuro considerar los siguientes puntos: tiempos de setup dependientes del servicio o tarea y generar una heurística correctiva posterior a la asignación, que permita reutilizar los horarios de forma que se cumplan lo más tempranamente posible, entre otros.

Agradecimientos: Este estudio tiene apoyo parcial de los proyectos: UDEC N°208.97011-1 y BASALCONICYT-FB0816.

Referencias

[1] Daganzo C.F., The crane scheduling problem. *Transportation Research Part B: Methodological*, 23 (2), 159-175. 1989.

[2] Hooker J. N., An Integrated Method for Planning and Scheduling to Minimize Tardiness. *Constraints*, 11 (2-3), 139-157. 2006.

[3] Kim K. H., Park Y., A crane scheduling method for port container terminals. *European Journal of Operational Research*, 156, 752-768. 2004.

[4] Kang J., Oh M., Ahn E., Ryu K., y Kim K., 2006. Planning for Intra-block Remarshalling in a Container Terminal. *Advances in Applied Artificial Intelligence*, 1211-1220. 2006.

[5] Lim A., Rodriguez B., Xiao F., y Ihue X., Crane scheduling with spatial constraints. *Naval Research Logistics*, 51, 386-406. 2004.

[6] Meisel F. y Bierwirth C., A unified approach for the evaluation of quay crane scheduling models and algorithms, *Computers & Operations Research*, vol. 38, nº. 3, 683-693. 2011.

[7] Ng W. C. 2005. Crane scheduling in container yards with inter-crane interference. European Journal of Operational Research, 164, 64-78.

[8] Pinedo M.L., 2008. Scheduling: Theory, Algorithms, and Systems. Springer, Third edition.

[9] Sammarra M., Cordeau J., Laporte G. y Monaco M., A tabu search heuristic for the quay crane scheduling problem. *Journal of Scheduling*, 10(4-5), 327-336. 2007.

[10] Wen C., Eksioglu S. D., Greenwood A., y Zhang S., Crane scheduling in a shipbuilding environment. *International Journal of Production Economics*, 124 (1), 40-50. 2010

Gestión de Capacidad en el Servicio de Urgencia en un Hospital Público

Carlos Reveco[*]
Richard Weber[**]

Resumen

Saturación, elevados tiempos de espera, y cargas horarias de trabajo desbalanceadas son problemas comunes en la mayoría de los hospitales públicos de Chile. A través de este trabajo se busca mejorar el flujo de pacientes en el servicio de urgencia de un hospital pediátrico por medio de la gestión y programación de capacidad del personal asistencial. Se han identificado dos etapas principales. En la primera, se debe estimar la demanda diaria del servicio de urgencia del hospital. La segunda parte requiere determinar los niveles de personal óptimos para todo tipo de trabajadores asistenciales necesitados en el servicio. Para esto se propone una estructura de turnos médicos, enfermeras y paramédicos óptima basada en programación lineal y con la información que se posee de la demanda futura. Los sistemas propuestos fueron implementados en un hospital público donde mostraron excelentes resultados.

Palabras Clave: *Pronóstico, Scheduling, Hospital, Programación lineal*

[*]Departamento de Ciencias de la Computación, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile

[**]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

1. Introducción

La adecuada planificación y programación de los recursos de un hospital influye directamente en la calidad de atención que se brinda a los pacientes. La ausencia de algún equipo, material o personal en el momento crítico, puede traer consecuencias tan graves que afecten la salud, integridad física e incluso la vida del paciente. Por otra parte un exceso de estos recursos genera costos innecesarios para el hospital. El desafío en la gestión de capacidad consiste entonces en cumplir con la demanda que a su vez es variable y no determinista.

A partir de la necesidad de conocer la cantidad de recursos que debe tener un hospital, en este trabajo se propone una metodología para encontrar la cantidad óptima requerida del principal recurso de cualquier centro prestador de servicios de salud, su recurso humano.

En el caso de los servicios hospitalarios la capacidad disponible de atención se ve determinada por las instalaciones físicas, tales como box de atención, y los recursos humanos, tales como médicos, paramédicos y enfermeras, que realizan diagnósticos y tratamientos. Esta capacidad se debe planificar para garantizar un buen nivel de servicio brindado y, a su vez, optimizar el uso de los recursos.

En este trabajo se presentan los dos enfoques principales del sistema que fue implementado en un hospital público y pediátrico en Santiago de Chile.

El primer enfoque es la predicción de la demanda donde se aplican modelos econométricos y de minería de datos. El segundo enfoque consiste en la programación de los recursos mediante la optimización.

En el capítulo se revisará brevemente la literatura relacionada con este trabajo. El capítulo se discutirá sobre el proceso que deben seguir los pacientes para conseguir una atención de urgencia.

El capítulo 3.3 detallará el enfoque utilizado para la predicción de la demanda y la categorización de los pacientes. El capítulo 4.2 se describirá el modelo de programación lineal planteado para optimizar los recursos en la sala de urgencia.

Finalmente en los capítulos 5.2 y 5.2 se mostrarán los resultados y las conclusiones, respectivamente.

2. Revisión Bibliográfica

Para poder hacer una correcta gestión de recursos, lo primero es contar con

una estimación de la demanda confiable. Muchos métodos diferentes se han propuesto para el pronóstico [16] [6]. Existen estudios que comparan diversos métodos de pronóstico en términos de precisión de los resultados. Uno de estos estudios, realizado por Adya [10], compara el pronóstico hecho con redes neuronales con otros métodos, concluyendo que las redes en general dan mejores resultados. Otros estudios [5] muestran, en casos prácticos, la superioridad de las máquinas vectoriales de soporte (Support Vector Machines o SVM) aplicadas a las series de tiempo sobre otras técnicas como modelos ARIMA o redes neuronales. También existen estudios publicados sobre el pronóstico en el área de la salud y, en particular, en la sala de urgencia. Algunos de estos se han centrado principalmente en la predicción del número de camas necesarias para satisfacer la demanda [14] [1]. Otros, más recientes, estudian el pronóstico de demanda pero basado en el enfoque tradicional de predicción con modelos ARIMA [9].

Por otra parte existe un gran número de publicaciones sobre la planificación de personal. La presente revisión se limitará a los trabajos más relevantes. Un tutorial introductorio a la programación de personal hecho por Blöchliger [8] que presenta los conceptos básicos del problema de programación y discute algunos aspectos de la programación del personal.

Existe poca evidencia de trabajos conducentes a determinar los niveles de personal médicos requeridos en los centros de salud. De hecho, los pocos estudios encontrados se basan en determinar estos niveles según estándares internacionales, los cuales básicamente indican que debe haber un determinado número de enfermeras y médicos por camas en cada área.

Otro estudio calcula la cantidad de personal que se requiere en un área determinada a través del número de pacientes que se encuentran en ésta. De esta forma, calcula la cantidad de horas por enfermera requeridas y por ende el total de este cargo necesario en esa área [11]. Este trabajo se centra en un análisis estadístico en lugar de un modelo de optimización; además sólo se enfoca en determinar el número de enfermeras y sólo en servicios diferentes a urgencias. Relacionado con la programación del personal, la literatura es mucho más extensa. De hecho, existe gran variedad de trabajos relacionados con la programación, incluso dentro del contexto de los sistemas de salud en el que se evidencian diversos tipos de problemas, como son: *nurse scheduling*, *patient scheduling* y *surgery scheduling*. Uno de los temas más desarrollados es la programación de enfermeras, donde se destaca el trabajo de Burke [2] en el cual se realiza una amplia recopilación de los principales trabajos relacionados con problemas de *nurse rostering*, los diferentes problemas que contiene, sus métodos de solución, la importancia de los mismos y sus respectivas fortalezas y debilidades. En el artículo de Warner [12] se elabora la programación de

las enfermeras teniendo en cuenta sus habilidades y minimizando el costo por ausencia de personal.

Existe también un trabajo interesante en un centro hospitalario de Colombia [15]; sin embargo, la gran diferencia con el presente trabajo radica en que aquel trabajo supone que la demanda de urgencia es fija, igual a la demanda histórica y constante dentro de la semana.

3. Proceso de atención en la sala de urgencia

El presente trabajo fue desarrollado en la urgencia de un hospital pediátrico público de la ciudad de Santiago. Es importante conocer el funcionamiento normal de la atención de urgencia ya que gran parte del proyecto se centrará en este servicio. Su flujo cuenta de varios pasos que se describirán a continuación.

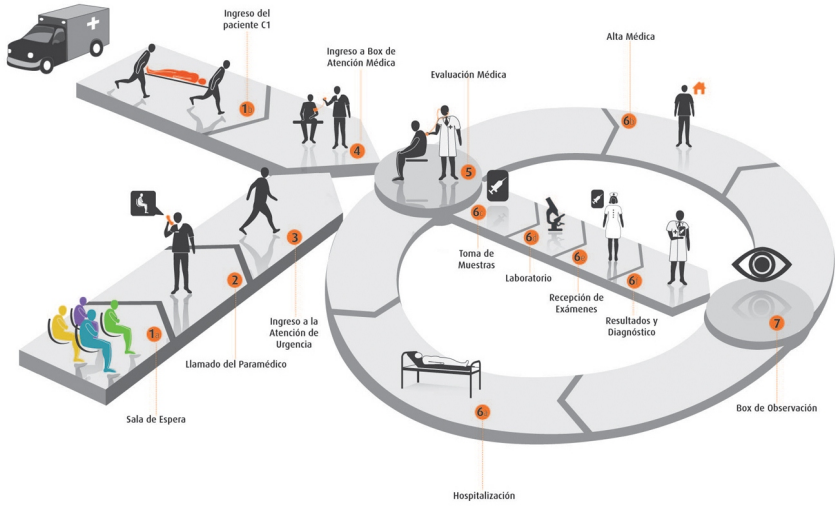


Figura 1: Proceso de atención de urgencia [20]

3.1. Selector de Demanda

Cada paciente que llega al servicio de urgencia tiene una categoría de urgencia que no es conocida a priori. En la primera etapa, la enfermera realiza un pre-diagnóstico a través de exámenes de rutina (temperatura, ritmo cardiaco, presión arterial) para poder categorizar al paciente. Al finalizar esta etapa, el paciente queda categorizado según su grado de urgencia. De acuerdo a lo definido por el Ministerio de Salud, existen cuatro categorías de pacientes:

- **C1: Paciente Grave:** Atención de emergencia, paciente con riego vital, evaluación y tratamiento inicial de inmediato.
- **C2: Paciente de alta complejidad:** Atención de urgencia, paciente agudo crítico, evaluación y tratamiento inicial inmediata.
- **C3: Paciente de mediana complejidad:** Atención de urgencia, paciente agudo no crítico, evaluación y tratamiento prioritario.
- **C4: Paciente no urgente:** Atención según disponibilidad de recursos, evaluación y tratamiento.

La urgencia viene condicionada por el tiempo hasta la atención definitiva mientras que la gravedad tiene más que ver con el pronóstico final. Por ejemplo, un cáncer puede ser más grave que una crisis de asma, pero ésta puede ser más urgente que el primero.

3.2. Ingreso y Atención

- **Ingreso de pacientes críticos:** Aquellos pacientes catalogados como Graves (C1) que vienen vía ambulancia, ingresan directamente a la atención médica suspendiendo el ritmo habitual de la urgencia.
- **Sala de espera:** Todo paciente no grave es atendido en admisión donde la secretaria llena una planilla con sus datos, pasa a clasificarse en el selector de demanda. El paciente deberá esperar a ser llamado para la atención médica.
- **Llamado del paramédico:** El paramédico le comunica al paciente que prontamente será atendido.
- **Ingreso a urgencias:** Tras ser llamado, debe ingresar al box que se le asigne para ser atendido.
- **Ingreso al box de atención:** Cabe destacar que para el hospital donde se desarrollo este trabajo hay dos tipos de box de atención; uno en pediatría, que atiende todas las enfermedades respiratorias, influenza, gripes, etc. y el segundo en traumatología donde llegan todas las atenciones por golpes, fracturas, contusiones, etc.
- **Evaluación Médica:** El médico de turno realizará la evaluación, para así determinar un posible diagnóstico.

3.3. Post-Evaluación

- **Toma de muestras:** La evaluación requiere que se le realicen exámenes médicos para determinar el diagnóstico y su futuro tratamiento.
- **Laboratorio:** Las muestras médicas de los exámenes deben ser evaluadas, lo cual toma aproximadamente 3 horas.
- **Recepción de exámenes:** La enfermera a cargo recibe los resultados y se lo comunica al médico.
- **Resultados y diagnóstico:** Con los resultados de los exámenes el médico determina el diagnóstico y su tratamiento el cual puede ser: Alta médica, box de observación u hospitalización.
  - **Alta médica:** Posterior al diagnóstico médico se prescribe el tratamiento correspondiente.
  - **Box de observación:** En el caso de requerir que el paciente espere de una forma supervisada, para ver la evolución de su problema de salud.
  - **Hospitalización:** Si el diagnóstico sugiere que sea internado, para realizar el servicio especializado en el hospital.

4. Pronóstico de Demanda

Para estimar la demanda de la sala de urgencia se proponen métodos basados en modelos de minería de datos. En este capítulo se presentan primero los modelos de pronóstico de la demanda agregada y luego una forma de separar el pronóstico por categorías.

4.1. Modelos de Pronóstico

Para poder predecir en forma efectiva, uno de los ingredientes clave es la calidad de la información y que las condiciones de funcionamiento del hospital y las condiciones ambientales permanezcan relativamente estables. Por ejemplo, si hay un terremoto o una pandemia, difícilmente el modelo podrá incorporar esa información y adaptar los pronósticos. Para ello se debería recalibrar los modelos para que realmente funcionen.

Los datos proporcionados por las bases de datos operacionales de urgencia en el hospital son de buena calidad ya que todo paciente debe ser registrado

en sus sistemas. Con una transformación, se pueden agregar los datos para llegar a la información de cantidad de pacientes.

La inspección visual de la demanda agregada revela un fuerte patrón estacional, como se muestra en la Figura 2. Además, se observa una leve baja en la demanda durante los meses de verano (enero - febrero) y una alta afluencia de pacientes durante los meses de la temporada de invierno (mayo - junio - julio). Esto se debe a que la contaminación del aire, el smog y las bajas temperaturas que conducen a enfermedades respiratorias, aumentan el número de atenciones, en especial en este hospital de niños.

Figura 2: Demanda histórica

Cuando los datos están desglosados por tipo de patología, es decir, separando por cada uno de los box de atención, pediatría y traumatología, se pueden ver grandes diferencias. La demanda en pediatría es mucho más volátil, ya que depende de factores tales como la temperatura y gripes estacionales, mientras que la demanda en el box de traumatología es más estable, como se muestra en las Figuras 3 y 4.

También es posible concluir de los datos que la demanda pediátrica agrupa un 70 % de los casos de urgencia y traumatología sólo un 30 %.

Para los pronósticos se aplicaron cuatro modelos diferentes: Regresión Lineal, Medias Móviles, Red Neuronal [17] y Support Vector Regression [7]. Para mayor detalle sobre la formulación de los modelos aquí planteados referirse al trabajo [13] y [3].

La validación de los modelos aquí desarrollados fue realizada utilizando la medida de comparación MAPE (*Mean Absolute Percentage Error*), definida como se muestra en la fórmula (1)

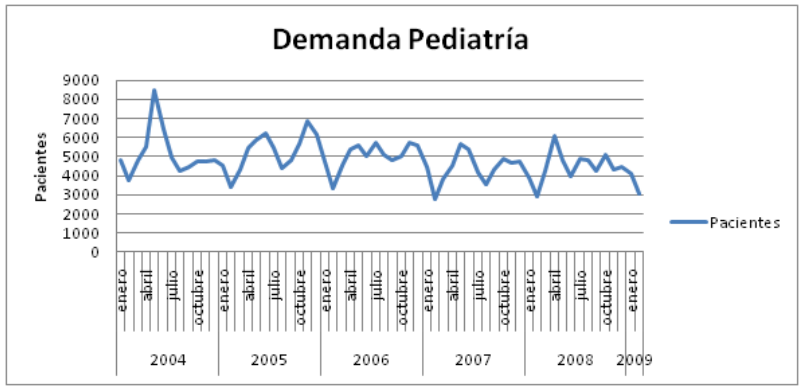


Figura 3: Demanda pediatría

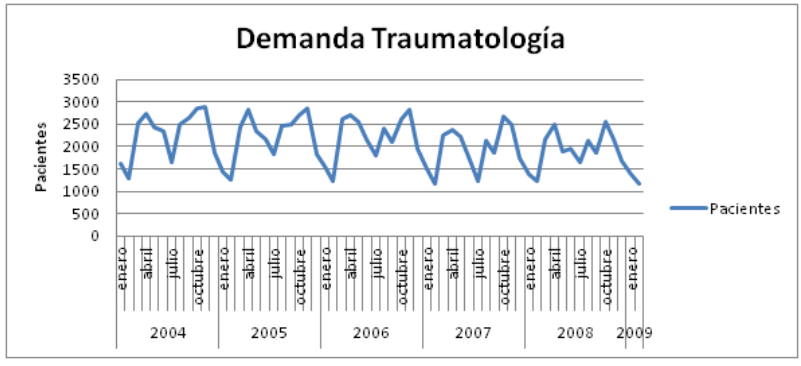


Figura 4: Demanda traumatología

$$[!h]MAPE = \frac{1}{N}\sum_{i=1}^{n}\frac{|Y_t - F_t|}{Y_t} \tag{1}$$

Donde Y_t es el valor real de la demanda para el período t y F_t el valor que se había pronosticado.

Los resultados obtenidos para cada uno de los modelos se muestra en la Tabla 1.

Una vez obtenidos estos resultados se segmentó la demanda mensual a una demanda semanal. Esto dado a que se logró evidenciar que cada semana dentro de un mes tiene un comportamiento similar. Entonces, tomando los pronósticos mensuales y asumiendo que la demanda es constante dentro de ese mes, se puede tener un pronóstico diario aplicando la distribución semanal y por turno, tal como muestran las Tablas 2, 3 y 4. Así se puede obtener el

	Regresión Lineal	Media Móvil	Red Neuronal	SVR
Pediatría	12,67 %	7,53 %	7,45 %	5,61 %
Traumatología	6,54 %	7,36 %	8,99 %	5,09 %

Tabla 1: Validación modelos de pronóstico

Día	Porcentaje
Lunes	16,7 %
Martes	14,6 %
Miércoles	14,1 %
Jueves	13,7 %
Viernes	13,6 %
Sábado	13,5 %
Domingo	13,9 %

Tabla 2: Porcentaje de atenciones diarias en pediatría

Día	Porcentaje
Lunes	16,5 %
Martes	16,7 %
Miércoles	16,1 %
Jueves	16,4 %
Viernes	14,3 %
Sábado	10,1 %
Domingo	9,9 %

Tabla 3: Porcentaje de atenciones diarias traumatología

dato aproximado de cuántos pacientes llegarán por cada turno durante un año corrido.

4.2. Pronóstico por Categoría

Uno de los temas importantes para el hospital, además de la cantidad total de los pacientes que llegan a la sala de urgencia, es la gravedad de cada uno de ellos, tal como se detalla en la sección 3.1 y se debe tener un estimado de cómo estos llegan dentro del día. Esto sirve para estar preparados ya que los pacientes más graves deben pasar más tiempo con los médicos que los pacientes menos graves. Teniendo esta caracterización se puede estimar la cantidad de recursos necesarios para una atención de calidad.

La gravedad de los pacientes puede ser obtenida en base a distribuciones

Turno	Horario	Porcentaje de atenciones pediatría	Porcentaje de traumatología
Día	08:00 – 19:59	73,46 %	75,85 %
Noche	20:00 – 07:59	26,54 %	24,15 %

Tabla 4: Distribución de atención por turno

históricas. Se observó que estas distribuciones tienen un comportamiento diferente a través de los distintos meses del año tal como se aprecia en la Figura 5, pero con un comportamiento relativamente similar a lo largo de los años.

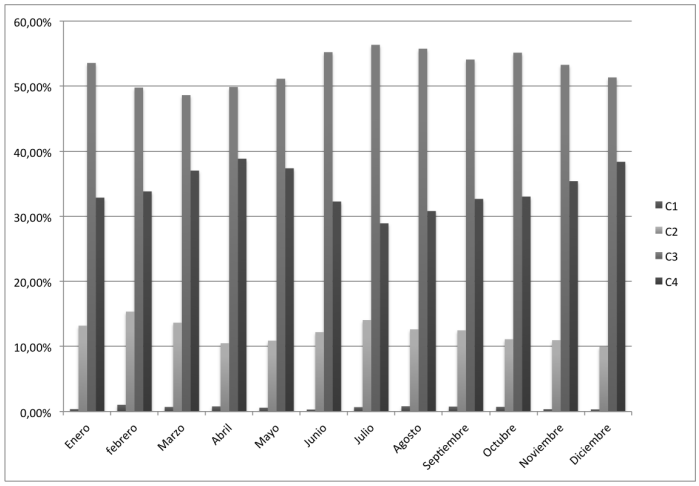


Figura 5: Distribución de categorías por mes

Los datos obtenidos por categoría para los diferentes meses se pueden ver de forma detallada en la Tabla 5

5. Modelo de Optimización para Determinar Personal en la Sala de Urgencia

El modelo busca determinar el número de médicos por turnos necesarios para cubrir de buena manera la demanda, pero no asigna específicamente a cada uno de los médicos. Esta asignación se hace actualmente en el hospital de forma manual.

La información de la cantidad de pacientes que llegan a la sala de urgencia se determina en base a los modelos de pronóstico descritos en el capítulo anterior. Estos pronósticos son desagregados en base a información histórica para separar la demanda diaria de pacientes y sus respectivas categorías.

A continuación se detallarán los parámetros y las variables de decisión para luego presentar el modelo de optimización.

	C1	C2	C3	C4
Enero	0,37 %	13,19 %	53,57 %	32,87 %
Febrero	1,03 %	15,35 %	49,79 %	33,83 %
Marzo	0,70 %	13,65 %	48,63 %	37,03 %
Abril	0,78 %	10,49 %	49,88 %	38,85 %
Mayo	0,58 %	10,89 %	51,14 %	37,38 %
Junio	0,30 %	12,20 %	55,23 %	32,27 %
Julio	0,67 %	14,05 %	56,36 %	28,92 %
Agosto	0,80 %	12,63 %	55,76 %	30,81 %
Septiembre	0,75 %	12,47 %	54,10 %	32,68 %
Octubre	0,71 %	11,10 %	55,15 %	33,03 %
Noviembre	0,35 %	10,96 %	53,28 %	35,40 %
Diciembre	0,32 %	9,95 %	51,35 %	38,38 %

Tabla 5: Distribución de categorías por mes

5.1. Definición de Parámetros y Variables de Decisión

En esta seccion se definen los conjuntos a utilizar, los parámetros y variables de decisión a ocupar por el modelo para determinar el personal para el hospital, considerando C categorías, D días, T turnos, A actividades y M meses como horizonte de planificación

c : Categoría de pacientes $c = 1 \dots C$
d : Días de la semana $d = 1 \dots D$
t : Turnos de trabajo $t = 1 \dots T$
a : Actividades a realizar $a = 1 \dots A$
m : Meses $m = 1 \dots M$

Los parámetros son:

$Minimo_{a,d,t,m}$	Número mínimo de trabajadores en la actividad a que debe haber en el turno t del día d en el mes m.
$Minutos_{a,c}$	Tiempo en minutos requerido para atender a un paciente de la categoría c en la actividad tipo a.
$Demanda_{a,c,d,t,m}$	Número de pacientes requiriendo atención en actividad tipo a de categoría c que ingresan en el día d en el turno t, durante el mes m.
$Jornadas_{a,m}$	Número de jornadas semanales disponibles para una actividad a en el mes m, en que cada jornada puede satisfacer un turno completo.
$q_{a,c}$	Número de personas destinadas a la actividad tipo a necesarias para atender a un paciente de categoría c.
o_t	Duración en minutos del turno t.
d_t	Tiempo en minutos destinado a descansos y suplementos, en cada turno t.
s_t	Minutos disponibles de trabajo en tiempo regular por turno t $s_t = o_t - d_t, \forall t = 1 \dots T$.
$r_{a,d,t,m}$	Minutos necesarios para cubrir actividad a en el día d en el turno t, en el mes m.

Variables de decisión:

$y_{a,d,t,m}$	Cantidad de médicos para la actividad a asignados para atender pacientes en el turno t, en día d y en el mes m

5.2. Modelo de Programación Lineal

El objetivo del modelo es minimizar la cantidad de trabajadores y por lo tanto el costo de personal en el horizonte de tiempo de planificación respetando las siguientes restricciones.

- Definir los minutos necesarios para la actividad a por turno t en base a las necesidades de Demanda, a los minutos requeridos por cada categoría c de pacientes para el día d en el mes m.

$$r_{a,d,t,m} = \sum_{c=1}^{C} (Demanda_{c,d,a,t,m} \cdot Minutos_{a,c} \cdot q_{a,c})$$
$$\forall d = 1 \dots D, \forall a = 1 \dots A, \forall t = 1 \dots T, \forall m = 1 \dots M, \forall c = 1 \dots C \qquad (2)$$

- Utilizar la cantidad de jornadas disponibles para asignar la cantidad de médicos a los diferentes turnos:

$$\sum_{d=1}^{D}\sum_{t=1}^{T} y_{a,d,t,m} \leq Jornadas_{a,m}$$
$$\forall a = 1 \ldots A, \forall m = 1 \ldots M \quad (3)$$

- Respetar el mínimo de trabajadores de cada tipo a en cada turno t de cada día d para cada mes m

$$y_{a,d,t,m} \geq Minimo_{a,d,t,m}$$
$$\forall d = 1 \ldots D, \forall a = 1 \ldots A, \forall t = 1 \ldots T, \forall m = 1 \ldots M \quad (4)$$

- Considerar el número de médicos en la actividad a en cada turno t de cada día d en el mes m para que sea capaz de atender emergencias (usualmente más de un médico):

$$y_{a,d,t,m} \geq q_{a,c}$$
$$\forall d = 1 \ldots D, \forall a = 1 \ldots A, \forall t = 1 \ldots T, \forall m = 1 \ldots M \quad (5)$$

- Naturaleza de las Variables

$$y_{a,d,t,m} \in \mathbb{N}$$
$$\forall d = 1 \ldots D, \forall a = 1 \ldots A, \forall t = 1 \ldots T, \forall m = 1 \ldots M \quad (6)$$

Minimizar la desviación de la capacidad requerida. Suponiendo que una desviación de la demanda se pondera de igual forma en ambos casos (demasiada capacidad o falta de capacidad). Se modela la función objetivo como valor absoluto de dicha desviación.

- Función Objetivo

$$Min|y_{a,d,t,m} * s_t - r_{a,d,t,m}|$$
$$\forall d = 1 \ldots D, \forall a = 1 \ldots A, \forall t = 1 \ldots T, \forall m = 1 \ldots M \quad (7)$$

Esta función no lineal se implemento como es común en estos casos, como suma de dos funciones excluyentes [4]. De la misma manera se puede diferenciar el costo de tener de tener demasiados recursos o falta de ellos por un ponderador.

Este modelo es genérico y flexible, ya que se puede adaptar a distintas situaciones con características similares al caso presentado. El modelo de programación lineal anteriormente descrito fue implementado en un software de optimización gratuito llamado ZIMPL [19] y resuelto a través del Solver no comercial SCIP [18]

6. Resultados

El modelo se probó con el pronóstico de septiembre del 2008 y se hizo una aproximación de turnos para el mes de septiembre del 2008. Los tiempos de atención para cada paciente según categoría fueron consensuados con el hospital tal como lo muestra la Tabla6.

C1	120 minutos
C2	60 minutos
C3	20 minutos
C4	10 minutos

Tabla 6: Tiempos de atención

También se consideró tiempos muertos en que el médico realiza otras actividades fuera de las estrictamente relacionadas con la atención de pacientes tal como se muestra en la Tabla 7

Administración	15 minutos cada 2 horas
Baño	15 Minutos cada 2 horas
Alimentación	45 minutos por turno
Docencia	20 minutos cada 2 horas de Lunes a Viernes de 14 a 20 hrs.

Tabla 7: Tiempo otras actividades

En el hospital debe haber un mínimo de un médico en cada especialidad en todo momento, por seguridad, si es que hay una emergencia. Se consideró que

Turno Mañana	8:00 - 19:59
Turno Tarde	20:00 - 7:59

Tabla 8: Estructura de atención.

para los casos C1 se necesitan dos médicos atiendan al paciente simultáneamente durante 120 minutos.

Los resultados del modelo se muestran gráficamente en las siguientes Figuras 6 y 7. Hay que destacar que en la actualidad hay tres médicos en pediatría y dos en traumatología para todo turno.

Observaciones

- El modelo es separable por mes.
- Los turnos se mantienen idénticos a los propuestos por el hospital, 2 turnos de 12 horas cada uno.

  Sin embargo, el modelo es fácilmente ampliable para probar distintos tipos de turnos, como turnos de 8 horas, y turnos de 1 hora que representa una granularidad máxima para ajustarse de forma perfecta a la demanda.
- Se asume que toda la demanda que llega en un turno debe ser atendida por ese mismo turno.

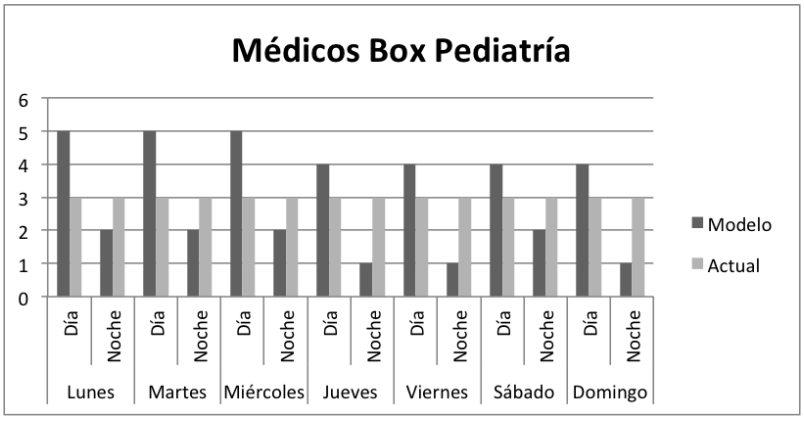


Figura 6: Médicos Box Pediatría

Dada la restricción de limitarse a distribuir las jornadas actuales de los médicos, se puede observar un mejor ajuste a la demanda. Para esta nueva distribución de turnos se estimo una mejora promedio de 10 minutos por paciente y quitando 4 turnos de traumatología.

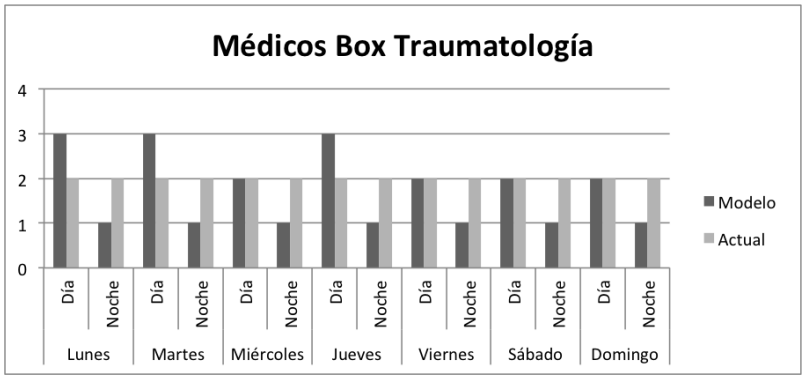


Figura 7: Box Traumatología

	Tiempo atención promedio [min]	Mejora %	Turnos necesarios	Capacidad Adicional
Situación base	81 min	-	70	-
Situación optimizada	71 min	12,3 %	66	-5,7 %
Óptimo calculado	45 min	43,2 %	76	8,6 %

Tabla 9: Resultados

Sin embargo eliminando la restricción de utilizar las jornadas actuales, el óptimo encontrado requiere 6 turnos adicionales lo que equivale a aumentar el total de jornadas en un 8,6 %. Para esta asignación de turnos se estimo que se reduce el tiempo promedio de espera por paciente en 36 minutos.

Lo anteriormente descrito se resume en la tabla 9

Si bien el modelo trata de minimizar la desviación de capacidad, en el momento del trabajo se operaba con un turnos diferentes a los calculados, lo cual hace que se trabajaba fuera del óptimo. Esto indica que la demanda no se estaba atendiendo de forma adecuada, es decir que había pacientes que debían esperar un tiempo considerable por una atención.

7. Conclusión y Trabajo Futuro

En este trabajo se aborda la problemática de la gestión de capacidad en un servio de urgencia de un hospital público. Esta unidad presenta características no determinísticas que hacen que la planificación no sea fácil. A través de modelos de predicción de demanda y un modelo de programación lineal es posible generar una metodología para realizar la asignación de turnos.

La predicción de la demanda es un proceso fundamental ya que es este el que alimenta con información confiable la gestión de recursos. Por lo tanto, una predicción confiable ayudará a aumentar la eficiencia del hospital y por sobre todo, el servicio de cara a los pacientes, quienes serán los más beneficiados gracias al mejor uso de recursos.

Como conclusiones generales del pronóstico de demanda podemos decir que la técnica Support Vector Regression (SVR) es la que genera mejores resultados; sin embargo, las redes neuronales también generan resultados bastante buenos.

El modelo de programación lineal aquí descrito incluye restricciones impuestas por el hospital y restricciones basadas en la capacidad de atención y la demanda esperada en base a los modelos de pronóstico. Este modelo indica que haciendo una mejor distribución de los turnos en urgencia, la espera de los pacientes se reducirá dramáticamente. Sin embargo, se debe aumentar la cantidad de recursos necesarios para esta distribución óptima, correspondiente a 6 turnos por semana. Este modelo es altamente flexible y reproducible con un alcance de anticipación hasta de un año, dada la capacidad de los modelos predictivos, por lo que es una excelente herramienta de gestión a mediano plazo.

Como trabajo futuro se puede realizar un modelo de optimización para la asignación de turnos, para los distintos tipos de trabajos, tarea que se hace actualmente en el hospital de forma manual.

Agradecimientos: El primer autor agradece al Magister en Negocios con Tecnologías de la Información (MBE) de la Universidad de Chile.

Este trabajo fue parcialmente financiado por el Instituto Sistemas Complejos de Ingeniería (ICM: P-05-004-F, CONICYT: FBO16).

Referencias

[1] Jones A., Joy M., and Pearson J. Forecasting demand of emergency care. *Health Care Management Science*, 2002.

[2] Burke, Causmaecker D., Berghe V., and Landeghem V. The state of the art of nurse rostering. *Journal of Scheduling*, 2004.

[3] Reveco C. *Pronóstico, Análisis y Gestión de demanda Hopitalaria*. Editoria Académica Española, 2011.

[4] Shanno D. and Weil R. 'linear'programming with absolute-value functionals. *Operations Research*, 1971.

[5] Velásquez J. D., Olaya Y., and Franco C. J. Predicción de series temporales usando máquinas vectoriales de soporte. *Revista chilena de ingeniería*, 2010.

[6] Box G. E., Jenkins G. M., and Reinsel G. C. *Time Series Analysis Forecasting and Control*. Prentice Hall, 1994.

[7] Drucker H, Burges C., Kaufman L., Smola A., and Vapnik V. Support vector regression machines. In *NIPS'96*, pages 155–161, 1996.

[8] Blöchliger I. Modeling staff scheduling problems a tutorial. *European Journal of Operational Research*, 2003.

[9] Schweigler L., Desmond J., McCarthy M., Bukowski K., Ionides E., and Younger J. Forecasting models of emergency department crowding. *Academic Emergency Medicine*, 2009.

[10] Adya M. and Collopy F. How effective are neural nets at forecasting and prediction? a review and evaluation. *Journal of Forecasting*, 1998.

[11] Warner M. Personnel staffing and scheduling. In *Patient Flow: Reducing Delay in Healthcare Delivery*, volume 91, pages 189–209. Springer US, 2006.

[12] Warner D. M. and Prawda J. A mathematical programming model for scheduling nursing personnel in a hospital. *Management Science*, 1972.

[13] Barros O., Weber R., Reveco C., Ferro E., and Julio C. Demand forecasting and capacity planning for hospitals. *Aceptado OR Spectrum*, 2010.

[14] Farmer R. and Emami J. Models for forecasting hospital bed requirements in the acute sector. *Journal of Epidemiology and Community Health*, 1990.

[15] Aguirre S., Amaya C., and Velasco N. Planeación y programación del personal del servicio de urgencias en un centro médico. *Los cuadernos de PYLO – Logística Hospitalaria*, 2008.

[16] Armstrong J. S. Principles of forecasting. 2001.

[17] McCulloch W. S. and Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysic*, 5:115–133, 1943.

[18] Achterberg T. Scip: Solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.

[19] Koch T. *Rapid Mathematical Prototyping*. PhD thesis, Technische Universität Berlin, 2004.

[20] López W., Aldea E., Riquelme G., Savoy S. Burgos J., and Velásquez M. Diseño del servicio de salud, caso hospital de quilpué, 2010. http://wiki.ead.pucv.cl/index.php/.

Caracterización de Contribuyentes que Presentan Facturas Falsas al SII Mediante Técnicas de Data Mining

Pamela Castellón [*]
Juan D. Velásquez[**]

Resumen

En este trabajo se entregan evidencias que es posible caracterizar y pronosticar a aquellos usuarios potenciales de facturas falsas en un año determinado, en función de la información de su pago de impuestos, el comportamiento histórico y sus características particulares, utilizando para ello distintas técnicas de Data Mining. En una primera instancia se aplican técnicas de SOM, Gas Neuronal y Árboles de Decisión para identificar aquellas variables que están relacionadas con un comportamiento de fraude y/o no fraude y detectar patrones de conducta asociada a esta problemática. Posteriormente se utilizan Redes Neuronales y Redes Bayesianas para establecer en qué medida se pueden predecir casos de fraude y no fraude con la información disponible. De esta forma se contribuye a identificar patrones de fraudes y generar conocimiento que pueda ser utilizado en la labor de fiscalización que realiza el Servicio de Impuestos Internos para detectar este tipo de delito tributario.

Palabras Clave: Facturas Falsas, Fraude Tributario, Data Mining, Clusterización, Predicción.

[*]Servicio de Impuestos Internos de Chile
[**]Departamento de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

1. Introducción

El fraude, en sus diversas manifestaciones, es un fenómeno del que no está libre ninguna sociedad moderna. Todas las instituciones, independiente de si son grandes o pequeñas, públicas o privadas, locales o multinacionales, se ven afectada por esta realidad que atenta gravemente contra los principios de solidaridad y de igualdad de los ciudadanos ante la Ley y pone en riesgo los negocios. De acuerdo a un estudio realizado por Ernst&Young en el año 2006 en el cual se encuestó a 150 empresas chilenas, medianas y grandes, un 41 % de ellas declaró haber sido víctima de algún tipo de fraude en los dos últimos años [8]. Esto plantea grandes desafíos en materia de detección y prevención, considerando que el fraude normalmente es mayor que lo declarado por las empresas, debido a que de alguna manera se resiente la imagen de la compañía y en muchos casos, incluso, hay empresas que no están en conocimiento de que han sido víctimas de un fraude.

La Evasión Tributaria y el Fraude Fiscal un tema que ha sido una constante preocupación de todas las administraciones tributarias, en especial de aquellas pertenecientes a países en vías de desarrollo[1]. Si bien es cierto, los impuestos no son la única fuente de financiamiento de un gobierno, es un hecho que éstos marcan una señal muy importante respecto al compromiso y la eficacia con que el Estado puede ejecutar sus funciones, y condicionar el acceso a otras fuentes de ingresos. En el caso de Chile, los ingresos tributarios proporcionan aproximadamente un 75 % de los recursos con que año a año el Estado sustenta sus gastos e inversiones, alcanzando durante el año 2010 un monto de $17,7 billones de pesos[2].

La utilización y venta de facturas falsas como mecanismo de evasión, es particularmente relevante, pues no sólo provoca una elusión de los impuestos, sino que en la mayoría de los casos implica un delito tributario. Por otra parte, junto a la generación de una merma en la recaudación, se producen efectos económicos negativos en el resto de las empresas, por el hecho de generar una competencia desleal frente a aquellas empresas que cumplen adecuadamente con sus obligaciones tributarias. Asimismo, se requiere que los recursos in-

[1] Habitualmente se habla de "elusión fiscal" cuando se hace referencias a conductas que, dentro de la Ley, evitan o reducen el pago de impuestos, mientras que la "evasión o fraude fiscal" supone un quebrantamiento de la legalidad para obtener para obtener esos mismos resultados.

[2] Información publicada en la Cuenta Pública SII 2010 de Marzo 2011, considerando los Ingresos Tributarios del Gobierno Central (sin incluir a Codelco, las Municipalidades y la Seguridad Social).

vertidos en fiscalización sean bien enfocados, detectando a aquellos de mayor riesgo de cumplimiento y no importunar ni desperdiciar tiempo y recursos en aquellos que si cumplen con sus obligaciones. Para ello, las técnicas de data mining ofrecen un gran potencial, ya que permiten extraer y generar conocimiento de grandes volúmenes de datos para caracterizar y detectar conductas fraudulentas y de incumplimiento para optimizar el uso de los recursos. Este artículo se organiza de la siguiente forma: en la sección 2 se describe la problemática e implicancias del uso de facturas falsas sobre la recaudación de los impuestos. La sección 3, describe la manera en que las técnicas de inteligencia artificial han facilitado la detección del fraude fiscal en otras administraciones tributarias. La sección 4 describe el acercamiento propuesto para caracterizar y detectar fraude en la emisión de facturas a través de las técnicas de data mining. La sección 5 presenta las principales conclusiones y las líneas de investigación futuras.

2. Necesidad de Detectar Fraude en un Institución Recaudadores de Impuestos

El Servicio de Impuestos Internos (SII) es la Institución responsable de administrar el sistema de tributos internos, facilitar y fiscalizar el cumplimiento tributario y propiciar la reducción de los costos de cumplimiento, en pos del desarrollo económico de Chile y de su gente. Para ello cuenta con 4.183 funcionarios, de los cuales el 31 % corresponde a fiscalizadores, quienes deben velar por el cumplimiento de 3.4 millones de contribuyentes, considerando los declarantes del Impuesto al Valor Agregado (IVA) y el Impuesto a la Renta.
Particularmente el IVA se ha convertido en un componente clave de la recaudación fiscal, representando durante el año 2010, el 47 % del total de los ingresos tributarios recaudados, por un monto de $8,3 billones de pesos [19]. Actualmente existen 708 mil contribuyentes que declaran IVA, de los cuales 28.000 están autorizados para emitir facturas electrónicas, lo cual ha ido aumentando progresivamente desde el año 2003, como parte de la política adoptada por el SII para modernizar su gestión y asegurar la autenticidad de los emisores de documentos tributarios. Del total de facturas emitidas, un 60 % se emite en formato papel y un 40 % en formato electrónico, generándose cerca de 400 millones de facturas al año.
El fenómeno de las facturas falsas respecto del IVA se explica por la mecánica de determinación del impuesto. Cuando una empresa recibe una factura falsa, aparenta con ello una compra que nunca existió, con lo que aumenta fraudu-

lentamente su crédito fiscal y disminuye su pago de IVA. Asimismo se produce una disminución del pago en el Impuesto a la Renta, debido al aumento de los costos y gastos declarados.

La falsedad del documento puede ser "material", si en él se han adulterado los elementos físicos que conforman la factura o "ideológica", cuando la materialidad del documento no está alterada, pero las operaciones que en ella se consignan son adulteradas o inexistentes. Ésta última es más difícil y compleja de detectar, ya que implica transacciones ficticias, en las cuales se requiere una auditoria para revisar los libros de compra y las rectificaciones o la realización de cruces de información con proveedores. Por otra parte, estos casos son más costosos para el Servicio, ya que requieren una mayor cantidad de tiempo destinado a la recopilación de antecedentes y pruebas, las cuales son más difíciles de encontrar.

Los casos más conocidos de falsedad material son la adulteración física del documento, la utilización de facturas colgadas en la que se falsifica una factura para suplantar a un contribuyente de buen comportamiento tributario, y el uso de doble juego de facturas, en la que se tiene dos facturas de igual numeración pero una de ellas ficticia y por un monto mayor. En el caso de la falsedad ideológica se encuentran las facturas utilizadas para registrar una operación inexistente o que adulteran el contenido de una operación existente. Adicionalmente existen otros delitos comúnmente relacionados, como la falsificación del inicio de actividades a través de palos blancos, con la única finalidad de adquirir facturas timbradas que posteriormente son vendidas a otros contribuyentes.

De acuerdo a un método de estimación de la evasión del IVA por concepto de facturas falsas y otros abultamientos de créditos, aplicado en el periodo 1990-2004 por el SII, la evasión por facturas falsas ha representado entre un 15 % y un 25 % de la evasión total del IVA, aumentando considerablemente en años de crisis económicas. Es así como en el año 1992, el porcentaje de participación aumentó a un 30 % y en la crisis del año 1998-1999 alcanza su punto máximo con un 38 % de participación, año en que alcanza una cifra cercana a los $317.000 millones de pesos. Esto adquiere relevancia producto que recientemente se produjo una crisis económica mundial que afectó a Chile a fines del 2008 y mediados del 2009, provocando un aumento de la tasa de evasión del IVA a un 18 %, por un monto evadido de $1,5 billones de pesos.

Asimismo, la detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera un importante costo administrativo para las áreas de fiscalización y jurídica. Durante el año 2010, el costo de recaudación de $100 fue de $0,91, es decir, aproximadamente un 1 % del valor recaudado. En el periodo 2001-2007 se han presentado

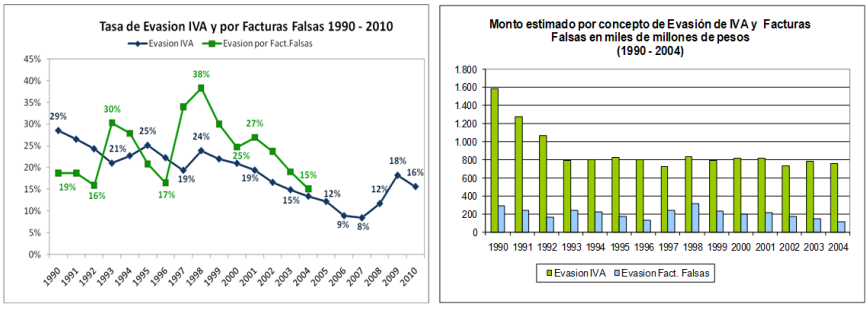


Figura 1: Tasa y Monto de Evasión en el IVA y por Facturas Falsas, Periodo 1990-2010 - Fuente: Subdirección de Estudios, SII

más de 2.300 querellas por facturas falsas y otros delitos de defensa judicial, las cuales involucraron a más de 4.000 querellados, por un monto de perjuicio fiscal cercano a los $274.130 millones de pesos.

Estadísticas SCE	2001	2002	2003	2004	2005	2006	2007	Acumulado
Cantidad de Querellas	171	394	358	407	451	306	243	2.330
Cantidad de Querellados	371	835	667	839	801	537	386	4.436
Monto Perjuicio Fiscal (MM$)	29.370	36.407	49.751	58.812	47.856	21.620	30.314	274.130
Casos SCE[3]	830	2.081	1.794	1.609	1.553	1.052	870	9.789

Tabla 1: Estadísticas de acciones legales relacionadas con facturas falsas 2001-2007 - Fuente: Cuenta Pública SII, 2005, 2006, 2007

El SII utiliza diversos métodos para seleccionar contribuyentes a ser controlados. En el caso de las fiscalizaciones masivas, los contribuyentes se determinan como resultado de un proceso de cruce de información de las declaraciones recibidas y otras fuentes de información, en la cual se detectan inconsistencias y diferencias tributarias. Las fiscalizaciones selectivas, en cambio, se generan en respuesta a determinadas figuras de evasión, ya sea a nivel nacional o local, utilizando para ello distintos ratios tributarios y condiciones, los cuales

utilizan información parcial del contribuyente. Para ello resulta fundamental, aprovechar la gran cantidad de información disponible en los sistemas respecto del comportamiento cada contribuyente en el tiempo.

3. Trabajos Relacionados

La mayor parte de las administraciones tributarias planifican su lucha contra el fraude fiscal. No obstante, existen importantes diferencias en los mecanismos, alcances, enfoque, contenido y énfasis puestos en dicha labor. Para detectar el fraude fiscal, las instituciones comenzaron aplicando auditorias de selección aleatoria o enfocándose en aquellos casos que no tuvieran fiscalizaciones en periodos anteriores recientes y seleccionando casos de acuerdo a la experiencia y conocimiento de los auditores [18]. Posteriormente, se desarrollan metodologías basadas en análisis estadísticos y en la construcción de ratios tributarios o financieros, lo cual evolucionó a la creación de sistemas basados en reglas y modelos de riesgo, que transforman la información tributaria en indicadores que permitan rankear a los contribuyentes por riesgo de cumplimiento. Durante los últimos años, las técnicas de Data Mining e Inteligencia Artificial, han sido incorporadas en las actividades de planificación de auditorías, principalmente para detectar patrones de fraude o de evasión, las cuales han sido utilizadas por las instituciones tributarias con fines específicos.
La Internal Revenue Service, institución a cargo de administrar los impuestos en Estados Unidos, ha utilizado técnicas de Data Mining con distintos fines, entre los que se encuentran la medición del riesgo de cumplimiento de los contribuyentes, la detección de la evasión tributaria y actividades financieras delictivas, la detección de fraude electrónico, la detección de abusos en impuesto de las viviendas, la detección de fraude en contribuyentes que reciben ingresos obtenidos por crédito fiscal y lavado de dinero [10]. Para ello ha utilizado modelos de regresión logística, árboles de decisión, redes neuronales, algoritmos de clustering y técnicas de visualización como Link Analysis, entre otros.
En la Administración Tributaria de Australia, el Çompliance Program" se basa en un modelo de riesgos, que utiliza estadísticas y Data Mining con el objetivo de realizar comparaciones, encontrar asociaciones y patrones mediante modelos de regresión logística, árboles de decisión y SVM cite18. Un caso de interés ha sido el enfoque utilizado por Denny, Williams y Christen [6] de descubrimiento de pequeños clusters o subpoblaciones inusuales, denominadas "Hot Spots", utilizando técnicas como el Self Organizing Map (SOM) para

explorar sus características, algoritmos de agrupación como k-means y representaciones visuales, que son fáciles de entender para usuarios no técnicos.
En el caso de Nueva Zelanda, el modelo existente asocia el grado de cumplimiento con la atención del control, el cual coincide con el utilizado por la administración australiana [18]. El Plan incluye un análisis del entorno económico, internacional, poblacional, de diversidad étnica y de estructura familiar. Por su parte, Canadá utiliza redes neuronales y árboles de decisión para distinguir las características de los contribuyentes que evaden o cometen fraude, en base a los resultados de auditorías pasadas, para detectar los patrones de incumplimiento o evasión [18].
A nivel latinoamericano, Perú fue uno de los primeros en aplicar estas técnicas para detectar evasión tributaria, incorporando al sistema de selección en la Aduana Marítima del Callao una herramienta de inteligencia artificial basada en redes neuronales [3]. Durante el año 2004, este modelo fue mejorado a través de la aplicación de reglas difusas y de asociación para el pre-procesamiento de las variables y árboles de clasificación y regresión (CART) para seleccionar las variables más relevantes. Por su parte, Brazil desarrolló el proyecto HARPIA (Risk Analysis and Applied Artificial Intelligence) de manera conjunta entre la Brazilian Federal Revenue y las universidades de ese país [7]. Este proyecto consiste en desarrollar un sistema de detección de puntos atípicos que ayude a los fiscalizadores a identificar operaciones sospechosas basado en la visualización gráfica de información de importaciones y exportaciones históricas, y un sistema de información de exportación de productos, apoyado en cadenas de markov, para ayudar a los importadores en el registro y clasificación de sus productos, evitar duplicidades y calcular para la probabilidad de que una cadena es válida en un determinado dominio.
En el caso de Chile, la primera experiencia fue desarrollada en el año 2007, utilizando SOM y K-means para segmentar contribuyentes de IVA de acuerdo a sus declaraciones de F29 y características particulares [13]. Posteriormente, siguiendo la tendencia internacional, en el año 2009 se construyen modelos de riesgos en distintas etapas del ciclo de vida del contribuyente, en los que se aplican técnicas de redes neuronales, árboles de decisión y regresión logística. Adicionalmente se desarrolla la primera experiencia para detectar potenciales usuarios de facturas falsas a través de redes neuronales artificiales y árboles de decisión, utilizando principalmente información de su declaración de IVA y Renta en micro y pequeñas empresas.

4. Aplicación de Data Mining para la Detección de Fraude en la Emisión de Facturas

A diferencia del estudio anterior desarrollado en el año 2009 relacionado con esta problemática, este trabajo busca complementar el uso de información de impuestos con variables adicionales relacionadas a su comportamiento histórico y su comportamiento en el año de análisis, así como incluir aspectos concernientes a sus relacionados directos, tales como mandatarios, socios y representantes legales. Por otra parte, se desarrolla un modelo para medianas y grandes empresas, en los que existe menor conocimiento de forma de operar respecto del uso de facturas falsas, debido a que tienen procedimientos más complejos de evasión.

4.1. Datos Utilizados

Para efectos de la caracterización se escoge el año 2006 como año de estudio. Si bien el peak de contribuyentes usuarios de facturas falsas detectados ocurre en el año 2002, se determina utilizar información más reciente, debido a que las dinámicas de evasión se van modificando en el tiempo, al igual que lo hicieron los formularios de pago de impuestos en ese periodo. Por otra parte, las auditorias se realizan hasta un periodo de 3 años atrás, lo que dificulta utilizar información más reciente, pues durante el año 2010 aún se estaban generando casos que podrían haber utilizado facturas falsas desde el año 2007 hacia adelante. De esta forma, el universo queda compuesto por todos aquellos contribuyentes que hayan presentado al menos una declaración de IVA entre el año 2005 y 2007, correspondiente a 582.161 empresas. Para caracterizar a los casos de fraude/no fraude se utiliza información de aquellas auditorias en las que existe certeza que se le revisaron sus facturas del año 2006, independiente del momento en el que fue realizada, generando un total de 1.692 empresas.

Contribuyentes del análisis	MI y PE	ME y GR	Total
Empresas activas en el periodo 2005-2007	558.319 (96 %)	23.842 (4 %)	582.161
Empresas auditadas por facturas en el 2006 con resultado de fraude o no fraude conocido	1.280 (76 %)	412 (24 %)	1.692

Tabla 2: Número de Contribuyentes Utilizados en el Análisis

Uno de los mayores inconvenientes para obtener la información de casos con fraude y no fraude se produce por la forma en la que se registra la información, pues se conoce la fecha de inicio y término de la auditoría, así como los periodos tributarios revisados y el resultado obtenido, pero la información de los periodos en los que ocurren las diferencias no está automatizada. Por lo tanto, para saber si la factura falsa detectada correspondía al año 2006 específicamente, hubo que revisar las anotaciones y comentarios efectuados por el auditor y las rectificatorias efectuadas en códigos relacionados con facturas de ese año.

Los casos de fraude y no fraude se categorizaron en tres tipos: "0" indica que el contribuyente fue auditado y no se encontraron facturas falsas en ninguno de los periodos revisados, "1" que indica que el contribuyente no utilizó facturas falsas en el año de análisis pero sí en otros periodos revisados (normalmente el año anterior o siguiente) y "2" que indica que el contribuyente utilizó facturas falsas en el año de estudio.

Para la construcción del vector de características se seleccionaron 20 códigos del Formulario de Pago Mensual de IVA (F29), 31 códigos del Formulario del Impuesto Anual de la Renta (F22) asociados a la generación de la base imponible de primera categoría y datos contables de la empresa, y 31 ratios tributarios que relacionan la información de IVA y Renta y la rentabilidad de la empresa con su liquidez, entre otros. Adicionalmente se generan 92 indicadores que pueden dar indicios de un buen o mal comportamiento en el tiempo, relacionados con su comportamiento histórico, el comportamiento de sus relacionados, sus características particulares e información generada en las distintas etapas del ciclo de vida, como se muestra en la Tabla Nº3.

4.2. Técnicas de Data Mining Implementadas

Para efecto de la caracterización e identificación de patrones, se aplican tres técnicas de data mining: el Self- Organizing Maps (SOM), el Gas Neuronal (NG) y Árboles de Decisión. Posteriormente para la predicción, se utiliza Redes Neuronales con Backpropagation y Redes Bayesianas, las que se describen a continuación:

- Self-Organizing Maps (SOM): es uno de los modelos de redes neuronales artificiales más utilizado para el análisis y visualización de datos de alta dimensión, basado en aprendizaje competitivo no supervisado. La red consiste en un conjunto de neuronas dispuestas en una grilla de dimensión a, normalmente rectangular, cilíndrica o toroidal, que genera un espacio de salida de dimensión d, con a$\leq$ d, sobre el cual se construyen relaciones de vecindad. Durante el entrenamiento de la red, las neuronas

Concepto	Tipo de Información
Pago de Impuestos	Declaraciones de IVA (F29), Declaración de Renta (F22), Ratios Tributarios de IVA/Renta
Características Propias	Edad, Antigüedad Empresa, Cobertura, Facturador electrónico, Contabilidad computacional, Actividades económicas, Cambio sujeto, Declara por internet, Tiene domicilio y sucursales propias
Comportamiento Histórico y en el año	Fiscalizaciones selectivas, Delitos Previos, Problemas con el domicilio, Inconcurrencias, Denuncias y Clausuras, Pérdidas de Rut, Destrucción de documentos, Deuda regularizada, Pérdida de Facturas, Facturas observadas y/o bloqueos, Marcas Preventivas.
Ciclo de Vida	Inicio de actividades, Verificación de actividades, Timbraje de documentos, Modificaciones de información, Términos de giro previos
Relacionados	Mandatarios, Representantes Legales, Socios, Familiares, Proveedores, Contadores, Sociedades y Representaciones (activos, antecedentes de delito, investigados, bloqueados)

Tabla 3: Tipo de Información utilizada para construir el vector de características

generan cierta actividad ante el estímulo de los datos de entrada, lo que permite determinar qué neuronas han aprendido a representar los patrones de la entrada, los cuales pueden ser agrupados dentro de una misma categoría o cluster, basándose en una medida de distancia, normalmente Euclideana. Esta herramienta usualmente es aplicada para clusterización y segmentación, generando grupos con objetos de comportamiento similar entre sí, pero diferentes a los objetos de otro grupo.

- Gas Neuronal (NG:Neural Gas): es un algoritmo relativamente nuevo de redes neuronales no supervisada, orientada a la cuantización vectorial de estructuras arbitrarias. La mayor diferencia con el SOM es que este método no define una grilla que impone relaciones topológicas entre unidades de la red y cada neurona puede moverse libremente a través del espacio de datos. Esta libertad permite al algoritmo una mejor capacidad para aproximar la distribución de los datos en el espacio de entrada, ya que las neuronas no están obligadas a tener que mantener ciertas relaciones de vecindad, sin embargo, requiere tener algunos antecedentes respecto del número de grupos que se espera obtener.

- Árboles de Clasificación: es uno de los métodos más utilizado para realizar clasificaciones, y se destaca por su sencillez y su aplicabilidad a diversas áreas e intereses. Básicamente el algoritmo consiste en formar todos los pares posibles y combinaciones de categorías, agrupando aquellas que se comportan homogéneamente con respecto a la variable respuesta en un grupo, manteniendo separadas las categorías que se comportan de forma heterogénea. Para cada posible par, se calcula el estadístico correspondiente a su cruce con la variable dependiente (estadístico chi-cuadrado en caso de campos de destino categóricos o estadístico F para salidas continuas). Para las categorías fusionadas se procede a realizar nuevas fusiones de los valores del pronosticador, pero esta vez con una categoría menos, El proceso se acaba cuando ya no pueden realizarse más fusiones porque los estadísticos entregan resultados significativos.

- Red Neuronal de Perceptrón Multicapa (MLP): es un modelo de red neuronal artificial de varias capas utilizado para la clasificación y agrupación, basado en la funcionalidad del cerebro humano a través de un conjunto de vértices interconectados. La red debe encontrar la relación existente entre los atributos de entrada y la salida deseada para cada caso. Esto lo realiza a través de un método de aprendizaje llamado "Backpropagation" o "Retropropagación del error", que minimiza el error de predicción mediante un ajuste a los pesos de la red. Este método posee dos etapas: en la primera se calculan las salidas basado en las entradas y los pesos asignados a la red inicial, para la cual se calcula el error de la predicción y en la segunda fase, se calcula el error hacia atrás a través de la red, desde las unidades de salida hacia las unidades de entrada. De esta forma se actualizan los pesos a través de un método de descenso por gradiente. Este proceso es iterativo, por lo que tras realizar varias veces el algoritmo, la red va convergiendo hacia un estado que permita clasificar todos los patrones que minimizan el error[4].

- Redes Bayesianas: son un grafo dirigido acíclico, utilizado para predecir la probabilidad de ocurrencia de diferentes resultados, sobre la base de un conjunto de hechos. La red consta de un conjunto de nodos que representan las variables del problema y de un conjunto de arcos dirigidos que conectan los nodos e indican una relación de dependencia existente entre los atributos de los datos observados. Las redes bayesianas describen la distribución de probabilidad que gobierna un conjunto de variables, especificando suposiciones de independencia condicional junto con probabilidades condicionales. Típicamente, este problema se divide en dos

[4]Normalmente se calcula el error cuadrático medio

partes: un aprendizaje estructural, que consiste en obtener la estructura de la red, y un aprendizaje paramétrico, en el que conocida la estructura del grafo, se obtienen las probabilidades correspondientes a cada nodo. Su principal ventaja es que permite obtener la probabilidad de ocurrencia de un determinado suceso en función de un conjunto de acciones, entregando una vista clara de las relaciones mediante un gráfico de red.

4.3. Pre Procesamiento de los Datos

La preparación de los datos es una parte fundamental del proceso KDD, ya que la información puede provenir de muchas fuentes, tener errores, ambigüedades o ser redundante, consumiendo gran parte del tiempo del proyecto. Por otra parte, los datos deben ser transformados de manera apropiada para realizar el análisis.

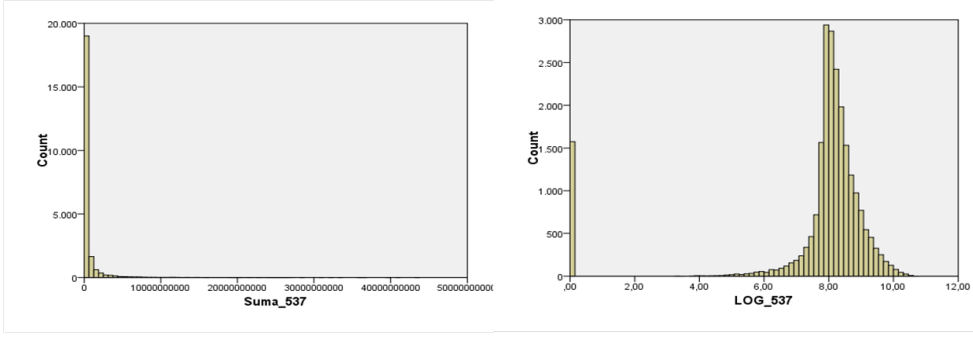
4.3.1. Limpieza

La calidad de los datos tiene una incidencia directa en los resultados, ya que si los datos no son de calidad, los resultados tampoco lo serán. Para lo anterior, se eliminan los puntos atípicos o outliers, utilizando como regla aquellos casos que superan la media más cinco veces la desviación estándar, considerando únicamente los casos con valor positivo de cada código. En la mayoría de las variables la distribución era decreciente, debido a que un gran porcentaje de contribuyentes paga montos bajos de impuestos, y sólo un pequeño grupo paga montos altos, por lo que la eliminación de datos se hizo de manera cuidadosa, considerando el juicio experto de los involucrados en el negocio, de manera de no eliminar casos que estuvieran correctos pero alejados del promedio. Lo mismo sucede con las variables de comportamiento, ya que constituyen conductas irregulares que sólo tiene un grupo pequeño de contribuyentes. Por lo tanto, al eliminar los casos con valores más altos, se elimina a aquellos contribuyentes que en general tienen un peor comportamiento, los cuales son el grupo de interés de este trabajo. Las variables de comportamiento, no tenían grandes inconsistencias debido a que fueron construidas en forma manual, sin embargo, se presentaban algunos problemas en los códigos del F29. Por ejemplo, se declaraban ventas con facturas pero no se indica una cantidad de facturas emitidas o viceversa. Dado que estos casos no eran muchos, se determina eliminarlos de la base. El mismo criterio se utilizó para el resto de los códigos de débitos y créditos.
Luego de quitar los outliers y los casos inconsistentes, el conjunto de datos final queda compuesto por 532.755 contribuyentes que son micro y pequeñas empresas, y 22.609 medianas y grandes empresas, eliminando un 4.6 % del

primer grupo y un 3.4 % del segundo.

4.3.2. Transformación y Normalización

Debido a que la declaración del pago de IVA se realiza mensualmente y la declaración de impuesto a la renta se realiza en forma anual, la primera transformación fue considerar el total anual, sumando los montos mensuales de cada código del F29 en el año para hacerlo comparable con la información de renta. Respecto de la completitud de datos nulos, la información de IVA es más completa que la de renta, debido a que los códigos del reverso del F22, sólo deben ser presentados por contribuyentes que llevan contabilidad completa. Por lo tanto, se utiliza información de débitos y créditos de IVA para completar datos de ingresos y costos del periodo, debido a la relación directa existente entre ambos.Para el resto de los campos de renta, se utiliza la mediana del código para contribuyentes del mismo tramo de ventas.Finalmente, producto de la distribución decreciente de las variables de impuesto, se aplica una transformación logarítmica para disminuir el efecto de los datos extremos como se muestra en la Figura N°2.



(a) Variable original x (b) Variable transformada Log(x+1)

Figura 2: Ejemplo de distribución original y transformada de códigos de impuestos.

Para evitar que las variables con un mayor rango de valores le quiten importancia a otras con un rango menor, se procede a normalizar las variables de manera que sean comparables la una con la otra, utilizando la normalización “Min-Max” en el rango [0,1]. Adicionalmente, previo a la selección de las variables de utilizar en los modelos, se procede a reducir las variables de comportamiento a través del Análisis de Componentes Principales (ACP). Como resultado se generan 15 componentes principales para el grupo de las micro y pequeñas empresas, que explican un 61,3 % de la varianza de los datos. Del

mismo modo, se generan 16 componentes principales para las medianas y grandes empresas, que explican un 59,9 % de la varianza de los datos, las que se presentan en la Tabla N°4.

Micro y Pequeñas Empresas	%	Medianas y Grandes Empresas	%
(1) Nivel de facturas timbradas en los últimos años	9,7	(1) Cobertura de la empresa	9,2
(2) Delitos e irregularidades de facturas previos	7,0	(2) Fiscalizaciones previas	6,2
(3) Fiscalizaciones previas con resultado positivo	5,6	(3) Nº Actividades económicas	5,5
(4) Frecuencia de Timbraje	5,1	(4) Nivel de formalidad de la empresa y antigüedad	4,2
(5) Participación en otras empresas	4,5	(5) Clausuras y denuncios históricos	3,8
(6) Problemas de localización	4,2	(6) Verificaciones de actividad	3,4
(7) Antigüedad	3,5	(7) Giros e inconcurrencias	3,2
(8) Clausuras y denuncias históricas	3.4	(8) Representantes legales	3,2
(9) Cobertura de la empresa	3,0	(9) Delitos de los relacionados	2,9
(10) Fiscalizaciones previas con resultado negativo	2,9	(10) Irregularidades de facturas y nivel de timbraje	2,8
(11) Verificaciones de actividad	2,6	(11) Rendimiento de fiscalizaciones previas	2,8
(12) Delitos de relacionados indirectos	2,6	(12) Irregularidades recientes	2,7
(13) Irregularidades previas (pérdida facturas)	2,5	(13) Cambio de sujeto	2,6
(14) Nivel de formalidad de la empresa	2,4	(14) Antecedentes de término de giro y no ubicado	2,6
(15) Delitos de relacionados directos	2,4	(15) Antecedentes de timbraje restringido	2,5
(16) Regularización de deudas y pérdidas de rut.	2,5		

Tabla 4: Conceptos asociados a cada Componente Principal y el porcentaje de la varianza explicada

Dado que nuestro interés era generar variables de comportamiento relacionadas al uso y venta de facturas falsas y no a otros comportamientos, se seleccionan sólo aquellas variables que tienen una correlación mediana-alta con

la variable de uso de facturas falsas en el año 2006, eliminando aquellas que tienen más de un 10 % de probabilidad que el coeficiente de pearson sea cero, exceptuando algunos códigos de interés como el total de débitos, total de créditos y pago de IVA, entre otros. Igualmente, se descartan aquellas variables que tienen un gran porcentaje de valores nulos. De esta forma se seleccionan 42 variables en el segmento micro y pequeñas y 36 variables medianas y grandes para el análisis. En el primer grupo, un 35 % de las variables corresponde a códigos de la declaración de IVA, un 35 % a códigos relacionados con renta y un 30 % a variables relacionadas al comportamiento. En el segundo grupo en cambio estos porcentajes varían a un 31 %, 38 % y 31 % respectivamente, con mayor preponderancia de variables relacionadas a la renta.

4.4. Modelamiento

Para efectos de caracterización e identificación de patrones, en una primera instancia se aplican las técnicas de data mining al universo de empresas, con el objetivo de identificar relaciones entre su pago de impuestos (IVA y Renta) y variables de comportamiento asociadas a la utilización de facturas falsas. Posteriormente se aplican técnicas de clasificación para aquellos casos en los que la condición de fraude y no fraude es conocido, de manera de identificar patrones específicos de este conjunto de contribuyentes. Finalmente se aplican herramientas de clasificación para predecir casos de fraude y no fraude con la información generada.

4.4.1. Caracterizando al Universo de Empresas

Inicialmente se aplica el método SOM al universo de contribuyentes, para identificar clusters o grupos de empresas de comportamiento similar. La hipótesis de trabajo suponía que al considerar sólo las variables de comportamiento relacionadas al uso de facturas falsas combinadas con variables de impuestos, era posible detectar grupos de contribuyentes que tienen un buen o mal comportamiento tributario y conocer cómo realizaban su pago de impuesto. Para ello se utiliza el paquete "som" de R, considerando una topología de red rectangular, con 3 neuronas de entrada y 24x24 neuronas de salida en el grupo de las micro y pequeñas empresas y 36x36 neuronas de salida en el grupo de las medianas y grandes empresas, con un número máximo de 100 iteraciones. En el primer grupo se considera una muestra de 100.000 empresas, debido a restricciones computacionales. En el caso de las micro y pequeñas empresas se generan 5 clusters, mientras que en las medianas y grandes se identifican 6 clusters, como se muestra en la Figura N°3.

Los clusters obtenidos en el primer grupo se diferencian principalmente

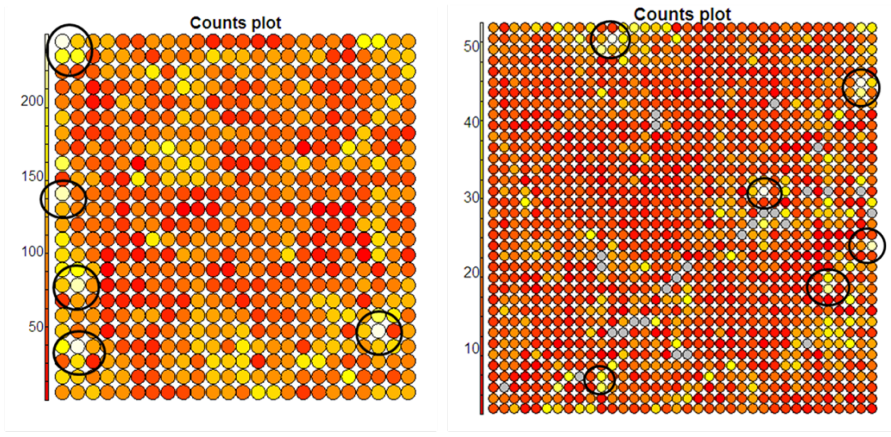


Figura 3: Mapa resultante aplicación SOM en MI y PE (izquierda) y ME y GR (derecha)

por la utilización de boletas y/o facturas, el nivel de pago de IVA, el nivel de costos declarados, el nivel de formalidad de la empresa y participación en otras empresas y algunos problemas de localización. Mientras que en las medianas y grandes, se diferencian por la utilización de boletas y/o facturas, niveles de uso de remanentes, notas de crédito y facturas de activo fijo, pasivos y activos, así como los resultados de fiscalizaciones previas y el nivel de formalidad de la empresa, como se indica en las Tablas N°5 y N°6.

Si bien se encontraron algunos patrones de comportamiento con éste método, estos no estaban relacionados específicamente a la utilización de facturas falsas, ya que los casos conocidos de fraude y no fraude se encontraban distribuidos en todo el mapa sin un patrón definido.

Posteriormente se aplica el Gas Neuronal, considerando el mismo número de clusters que el Mapa de Kohonen, utilizando el paquete "Clust" de R, el cual genera una matriz con las características de los centroides de cada variable y un vector de clasificación que señala el grupo al que pertenece cada contribuyente. En este caso, los grupos generados también se encuentran influenciados por el pago de impuestos, aunque con mayores diferencias en términos de comportamiento. Esto, permite diferenciar cuáles grupos tienen mejor y peor comportamiento, y relacionarlo con su pago de impuesto, aunque no necesariamente los casos de facturas falsas se encontraban en un mismo grupo.

De acuerdo a esto, se identificaron los siguientes patrones asociados a un mal y buen comportamiento, considerando los puntos comunes obtenidos en ambos métodos.

Cluster 1	No utiliza boletas y tiene nivel intermedio de uso de facturas, nivel alto de pago de IVA y costos altos. Con algunos problemas de localización, mayor nivel de participación en otras empresas y formalización de la contabilidad.
Cluster 2	No utiliza boletas y tiene nivel intermedio de uso de facturas, nivel intermedio-alto de pago de IVA y costos mínimos. No tiene problemas de localización reciente y presenta bajo nivel de formalidad y participación en otras empresas.
Cluster 3	No utiliza boletas y tiene poco uso de facturas, no genera IVA, aunque tiene nivel intermedio de pago, probablemente por los PPMs. Declara costos mínimos. No tiene problemas de localización reciente y presenta bajo nivel de formalidad.
Cluster 4	No utiliza boletas y tiene poco uso de facturas, no genera IVA, aunque tiene nivel intermedio de pago, probablemente por los PPMs. Declara niveles altos de costos y problemas de localización.
Cluster 5	Tiene niveles altos de débitos con boletas, nivel intermedio de uso de facturas y pago de IVA, y costos altos. Relativamente joven con algunos problemas de localización y nivel intermedio de formalización.

Tabla 5: Clusters resultantes aplicación SOM en MI y PE

4.4.2. Caracterizando a los Casos con Fraude y Sin Fraude

Si bien las dos técnicas anteriores implementadas permiten caracterizar al universo de contribuyentes e identificar algunos patrones diferenciadores, considerando aquellas variables más relacionadas con el uso de facturas falsas. Éstas tienden a darle mayor importancia al pago de impuestos que a las variables de comportamiento, creando grupos que se diferencian en el tipo de operación (ventas con facturas y/o boletas), el nivel de actividad (alto-bajo nivel de ventas, costos) y pago de impuestos (alto-bajo), debido a la mayor variabilidad de estas variables en comparación a las de comportamiento.
Por otra parte, al analizar la distribución de cada variable, se observa que los casos con fraude normalmente se encuentran en los casos extremos de cada una de ellas. Por este motivo se determina aplicar árboles de decisión al conjunto de datos con resultado de auditoría conocido, ya que permite identificar el punto de corte de cada variable frente al cual se produce un cambio de comportamiento, considerar casos extremos y generar reglas que pueden ser validadas e implementadas.

Cluster 1	No utiliza boletas. Tiene nivel intermedio de remanentes y costos bajos. Presenta monto alto de créditos por factura de activo. Con un nivel alto de formalidad.
Cluster 2	No utiliza boletas. Tiene nivel intermedio de remanentes y pocas fiscalizaciones previas. Nivel intermedio de formalidad.
Cluster 3	No utiliza boletas. Tiene nivel alto de remanentes, pasivos y activos. Tiene bajo porcentaje de crédito asociado a facturas. Nivel alto de formalidad.
Cluster 4	Nivel alto de uso de boletas. Tiene nivel intermedio de remanentes y de notas de crédito. Nivel alto de formalidad.
Cluster 5	Nivel alto de uso de boletas. Tiene pocos remanentes y nivel bajo de formalidad. Pocas fiscalizaciones previas.
Cluster 6	Nivel alto de uso de boletas. Tiene pocos remanentes y nivel alto de formalidad. Tiene nivel intermedio de uso de notas de crédito.

Tabla 6: Clusters resultantes aplicación SOM en ME y GR

El tipo de árbol utilizado es el CHAID (Chi-square automatic interaction detection), el cual permite clasificaciones no binarias y generar un número distinto de ramas a partir de un nodo considerando tanto variables continuas como categóricas. Un punto a considerar de éste método es que se requiere disponer de tamaños de muestra significativos, ya que al dividirse en múltiplos grupos, cabe el riesgo de encontrar grupos vacios o poco representativos si no se dispone de suficientes casos en cada combinación de categorías. Adicionalmente se evalúa el método del CHAID exhaustivo, el cual es una modificación del algoritmo tradicional, que busca hacer frente algunas debilidades del CHAID tradicional.
Se realizan varios experimentos que consideran distinto número de variables y tipos de salidas (categóricas y numéricas) para identificar si se producen diferencias entre una formato de salida y otro.

Finalmente esta técnica resultó ser altamente efectiva para encontrar patrones diferenciadores entre fraude y no fraude, ya que los nodos finales estaban compuestos mayoritariamente por casos de un solo tipo, o en su defecto combinado con casos con valor de salida “1”, los cuales se aproximan más al comportamiento de los casos con fraude “2”.
Como se indica en la Tabla N°8 el número de nodos finales fue similar en ambos experimentos realizados en cada grupo, obteniéndose 33 y 36 nodos en el segmento de las micro y pequeñas empresas y 22 y 24 nodos en el segmento

Buen Comportamiento MI y PE	Declaran montos más altos de débitos (emite más boletas) y pagan más IVA. Declaran bajos niveles de créditos y de remanentes, mayor relación ingresos/costos y costos/activos. Tienen mayor cantidad de facturas timbradas y frecuencia de timbraje, menor cantidad de delitos e irregularidades previas y delitos de los relacionados indirectos. Registran pocas verificaciones de actividad.
Buen Comportamiento ME y GR	Declaran mayor nivel de costos y gastos y mayor nivel de activos y pasivos. Tienen montos más altos de créditos y remanentes. Registran un mayor nivel de formalización de su contabilidad y mayor cobertura, mayor número de representantes legales y cantidad de fiscalizaciones previas.
Mal Comportamiento MI y PE	Declaran niveles bajos de pago de IVA y una relación débito/crédito baja. Registran una mayor cantidad de créditos y acumulación de remanentes. Tienen un nivel más bajo del ratio ingresos/activo, mayor cantidad de fiscalizaciones previas con resultado positivo y un menor nivel de facturas timbradas. Registran varias verificaciones de actividad.
Mal Comportamiento ME y GR	Declaran mayores costos y remuneraciones respecto de sus activos, menor nivel de pasivos y mayor cantidad de porcentaje de débitos con boleta, aunque con un número menor de boletas. Registran mayor cantidad de anotaciones de timbraje restringido, términos de giro previos y antecedentes de no ubicado. Tienen mayor cantidad de denuncias y clausuras históricas, menor cantidad de fiscalizaciones previas y cobertura, así como un menor nivel de formalización de su contabilidad y antigüedad.

Tabla 7: Caracterización de grupos con buen y mal comportamiento

de las medianas y grandes.

A modo de ejemplo se presenta un extracto del resultado de la aplicación del experimento N°1, en el cual se identifican patrones bastante claros asociados a fraude y no fraude, debido a la preponderancia de nodos finales con casos de fraude y no fraude. Como se indica en la Figura N°4, los factores que tienen mayor incidencia fueron el resultado de las fiscalizaciones previas (ACP10) y

Exp. Nº	Segmento	Método	Nº Variables	Tipo de salida	Nº Niveles	Nº Nodos finales
1	Micro y Peq.	Árbol CHAID	30	Categórica	6	33
2	Micro y Peq.	Árbol CHAID	30	Numérica	5	36
3	Med. y Grandes	Árbol CHAID	38	Numérica	4	22
4	Med. y Grandes	Árbol CHAID	24	Numérica	6	24

Tabla 8: Caracterización de grupos con buen y mal comportamiento, según el gas neuronal

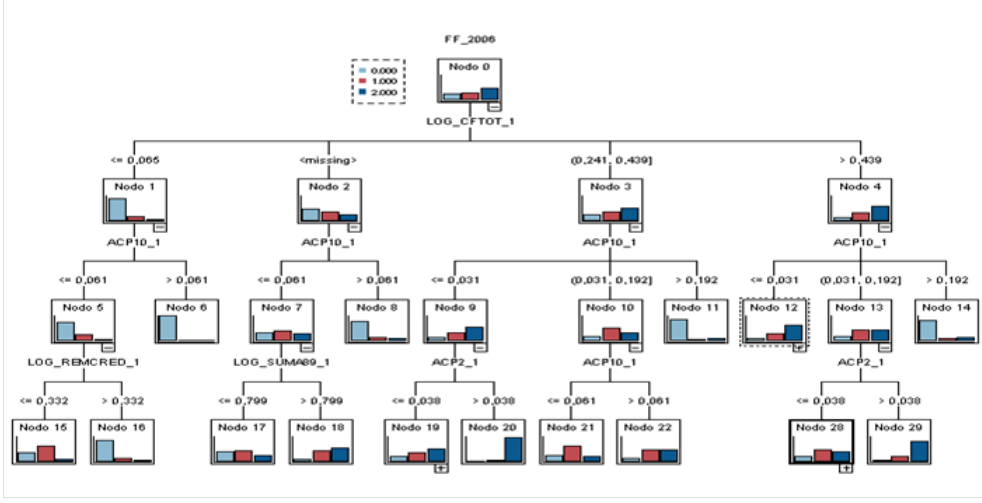


Figura 4: Clasificación resultante de la aplicación del árbol CHAID – Experimento N°1

el porcentaje de las compras sustentado en facturas (CFTOT). Esto indica que aquellos que han sido más veces fiscalizados en el pasado y no se les ha encontrado nada y sus compras no se basan principalmente en facturas, tienen menos probabilidad de utilizar facturas falsas, que aquellos que mayoritariamente registran compras con facturas y tienen fiscalizaciones productivas en el pasado. De hecho, estas dos variables por sí solas, determinan varios nodos finales con preponderancia de casos sin fraude.

Adicionalmente, la variable que indica una mayor preponderancia de delitos e irregularidades asociadas a facturas históricas combinado con la frecuencia de timbraje, genera nodos finales con preponderancia de casos con facturas falsas. Particularmente el nodo 12 que contiene casi la mitad de los casos (46 %) se descompone en varias ramas en función del valor que toma el crédito promedio por factura emitida (mientras mayor sea este indicador, mayor posibilidad

hay de que cometa fraude). De igual manera, la preponderancia de casos con fraude en cada rama depende del número de facturas emitidas, el IVA pagado, el total de débitos por boletas, la relación entre costos y activos y el nivel de participación en otras empresas.
Cómo se señala en la Figura N°5, las variables más relevantes para distinguir casos de fraude en las micro y pequeñas empresas fueron el resultado de las fiscalizaciones previas, el Total de IVA determinado, el porcentaje de crédito sustentado en facturas, la relación entre remanentes y créditos, el total de débitos por boletas y la relación entre facturas timbradas y emitidas. Mientras que en las medianas y grandes las variables corresponden a total de remanente, porcentaje de crédito respaldado en facturas, el número de representantes legales, nivel de formalización de la contabilidad, la relación entre remuneraciones y activos, entre otros.

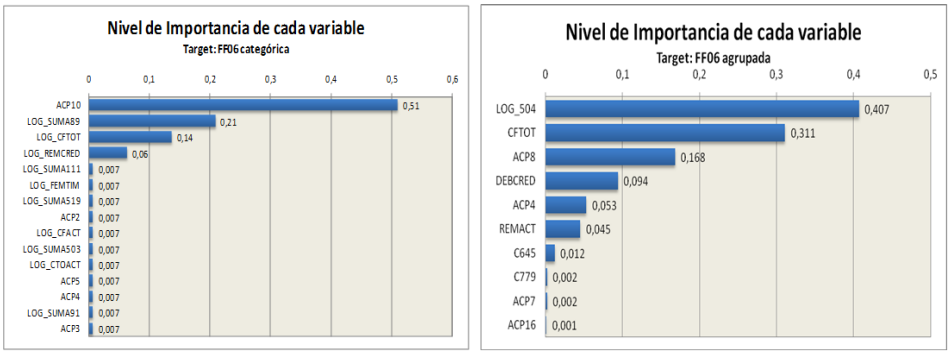


Figura 5: Nivel de importancia de las variables en cada grupo de acuerdo a la red neuronal

Considerando los patrones y reglas que se repiten en cada rama del árbol para diferenciar entre casos de fraude y no fraude, en la Tabla N°9 se presenta un extracto de los comportamientos asociados a cada uno de ellos en cada segmento, que resume las variables principales consideradas y las relaciones que generan nodos con y sin utilización de facturas falsas en el año de estudio.

4.4.3. Predicción del Fraude

Para la predicción, se aplicaron redes neuronales artificiales y redes bayesianas. En ambos procesos para evitar el sobreajuste de la red, los datos se dividen en dos conjuntos: uno de entrenamiento y uno de testeo, utilizando la regla 70/30. Por otra parte, ambos métodos fueron implementados utilizando la herramienta tecnológica clementine del SPSS.
Uno de las complejidades de las redes neuronales, es determinar el número de capas y nodos ocultos, así como la cantidad de épocas o iteraciones. Para

Comportamiento Asociado a Fraude	MI y PE Registran menor porcentaje de créditos asociados a facturas y más fiscalizaciones previas con resultado negativo. Emiten menor cantidad de facturas emitidas y un valor más bajo del indicador facturas emitidas/facturas timbradas. Registran un mayor monto del indicador remanentes/crédito promedio.
Comportamiento Asociado a Fraude	ME y GR Registran menor porcentaje de crédito asociado a facturas. Declaran un monto mayor de remanente acumulado del periodo anterior. Tienen valores bajos del indicador costos/activos. Registran menor cantidad de irregularidades previas asociadas a facturas y de timbraje.
Comportamiento Asociado a No Fraude	MI y PE Tienen mayor porcentaje de créditos asociados a facturas y débitos con boletas. Tienen valor alto del indicador costos/activos. Emiten una mayor cantidad de facturas y tienen valor alto del indicador facturas emitidas/facturas timbradas. Tienen montos altos de IVA determinado. Registran menos fiscalizaciones previos con resultado negativo y más fiscalizaciones previas con resultado positivo. Tienen más antecedentes de delitos e irregularidades históricas asociadas a facturas y mayor frecuencia de timbraje en los últimos dos años.
Comportamiento Asociado a No fraude	ME y GR Tienen mayor porcentaje de créditos asociados a facturas. Declaran monto menor de remanente acumulado en el mes anterior y tienen valores altos del indicador costos/activos. Tienen mayor nivel de informalidad en su contabilidad y son de menor antigüedad. Registran mayor número de actividades económicas activas e irregularidades previas asociadas a facturas y timbraje. Tienen mayor cantidad de giros e inconcurrencias a notificaciones.

Tabla 9: Caracterización de casos con y sin fraude según árbol CHAID

determinar tales parámetros se consideraron distintos números de ciclos y nodos en las capas ocultas, de manera de establecer a través de ensayo y error los valores más adecuados. Para las iteraciones se utilizaron los valores: 1.000, 5.000, 10.000 y 20.000. En el caso de los nodos se utiliza el número que el software calcula por defecto en función de los datos del modelo y otra correspondiente a la mitad del número de nodos de entrada, es decir, 3 y 20 nodos respectivamente.

En el caso de las redes bayesianas se evalúan dos métodos para construir la red: el algoritmo TAN y el algoritmo de estimación de Markov-Blanket disponibles

en el software clementine del SPSS. Adicionalmente se utiliza un preprocesamiento previo de las variables para identificar cuáles son las variables más relevantes y mejorar el tiempo de procesamiento y rendimiento del algoritmo. De igual forma se utiliza un test de independencia de máxima verosimilitud y chi-cuadrado para el aprendizaje paramétrico.
Los resultados de los experimentos se presentan en la Tabla N°10, el que contiene los siguientes indicadores obtenidos en el grupo de testeo: (1) Sensibilidad: Indica la proporción de casos con fraude clasificados en forma correcta, (2) Especificidad: Indica la proporción de casos sin fraude en los que la clasificación fue correcta, (3) Concordancia: Indica la proporción de casos con y sin fraude en los que la clasificación fue correcta y (4) Tasa de error: Indica la proporción de casos con y sin fraude que fueron asignados a una clase incorrecta.

Exp. N°	Segmento	Método	Sensitivi -dad (1)	Especifici -dad (2)	Concordan -cia (3)	Tasa Error (4)
1	Micro y Peq.	Red Neuronal	92.6 %	72.9 %	87.2 %	12.8 %
2	Micro y Peq.	Red Bayesiana	82.3 %	64.1 %	77.9 %	22.1 %
3	Med. y Grandes	Red Neuronal	84.3 %	52.2 %	65.8 %	34.2 %
4	Med. y Grandes	Red Bayesiana	73.3 %	66.7 %	70.3 %	29.7 %

Tabla 10: Experimentos realizados para predecir los casos con fraude por facturas falsas

En ambos segmentos, los mejores resultados de predicción de casos con facturas falsas se obtuvieron con la técnica de red neuronal. En el grupo de las micro y pequeñas empresas, el experimento 1 arrojó que en un 92,6 % los casos con fraude fueron asignados a la clase correcta, mientras que en el grupo de las medianas y grandes empresas la proporción de casos con fraude correctamente asignada fue de 84.3 %. Por otra parte, el poder de generalización del modelo fue bastante bueno, ya que los resultados del testeo fueron similares a los obtenidos en el entrenamiento de la red, cuya predicción fue casos con y sin fraude fue de 93.7 % y 89.6 % respectivamente.

La red neuronal generada para las micro y pequeñas empresas, indica una preponderancia de variables asociadas al pago de IVA y al comportamiento, y en menor medida, a variables relacionas a la renta. Las más relevantes corresponden a los antecedentes obtenidos de la verificación de actividades, la

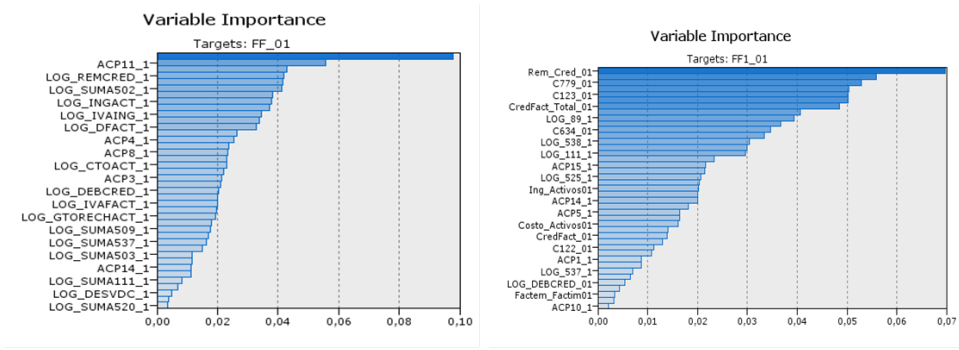


Figura 6: Nivel de importancia de las variables en cada grupo de acuerdo a la red neuronal

relación entre remanentes y créditos, el total de débitos por facturas emitidas, la relación entre ingresos del giro y los activos y la relación entre el IVA pagado y el Ingreso declarado. En el caso de las medianas y grandes empresas, las variables más relevantes corresponden a la relación entre remanentes y créditos, las cuentas por pagar a empresas relacionadas, el total de pasivos, la proporción de créditos asociado a facturas y el IVA determinado en el periodo.

5. Conclusión y Trabajo Futuro

La utilización y venta de facturas falsas tiene un impacto significativo en la recaudación que percibe el Estado para financiar sus proyectos. La detección, investigación, sanción y cobro de los impuestos adeudados, como consecuencia del uso de estos documentos, genera además un importante costo administrativo para el SII, lo que da cuenta de la relevancia que tiene focalizar los esfuerzos en la detección de casos de evasión y fraude fiscal.

Los métodos de clusterización y clasificación utilizados para caracterizar a los contribuyentes que tienen buen o mal comportamiento tributario asociado a la utilización de facturas falsas, demuestran que es posible identificar algunas características diferenciadoras entre un grupo y otro, las cuales hacen sentido con lo que sucede en la realidad. Particularmente el método de gas neuronal arrojó que era posible determinar algunas variables relevantes para diferenciar entre un buen o mal comportamiento, los que no necesariamente se asocian a la utilización y venta de facturas falsas. El método de kohonen, en cambio, no permitió obtener patrones de comportamiento relacionados con la utilización de facturas falsas, sino más bien, se detectaron clusters en relación al pago

de impuestos, en la que las variables con mayor cantidad de ceros y varianza resultaron ser las que más impacto tuvieron en la conformación de los grupos. Los árboles de decisión aplicados a los casos en el que el resultado de fraude y no fraude era conocido resultó ser una buena técnica para detectar variables que permiten distinguir entre casos de fraude y no fraude. Esto debido que al analizar la distribución de las variables en cada grupo, se observa que los casos con fraude tendían a tomar valores más extremos de las variables, por lo que era posible distinguir rangos a partir de los cuales, existe una probabilidad de tener o no tener fraude. Por otro lado, los resultados obtenidos fueron coherentes con lo observado en la realidad, de acuerdo a la vista experta.

Es así como en el caso de las micro y pequeñas empresas las variables que permitían distinguir entre fraude y no fraude se relacionaban principalmente con el porcentaje de créditos generado por facturas respecto del crédito total y las fiscalizaciones previas con resultado negativo. En la medida que el contribuyente fue fiscalizado más veces en el pasado y no se encontró nada, es más probable que no tenga fraude en el futuro. Por otro lado, mientras su crédito esté más asociado a otros ítemes distintos a las facturas (activo fijo u otros), es menos probable que utilice facturas para respaldar sus créditos. Otras variables relevantes fueron la cantidad de facturas emitidas en el año y su relación con las facturas timbradas en los últimos dos años, el monto de IVA total declarado, la relación entre remanentes y créditos promedio, las fiscalizaciones previas con resultado positivo y los delitos e irregularidades históricos asociadas a facturas. Mientras que en las medianas y grandes empresas, las variables más relevantes fueron la cantidad de remanente acumulado en los periodos anteriores, el porcentaje de crédito asociado a facturas, la relación entre costos y activos, el nivel de informalidad en su contabilidad y la antigüedad, así como la cantidad de irregularidades previas asociadas a facturas y la cantidad de giros e inconsciencias históricas.

En relación a los modelos predictivos, los que tuvieron mejor desempeño fueron los modelos de red neuronal de perceptrón multicapa, que para efectos del estudio contaban con una capa de entrada que contenía las variables explicativas, una capa intermedia de procesamiento y una capa de salida. En el caso de las micro y pequeñas empresas el porcentaje de casos con fraude asignado correctamente fue un 92 %, mientras que en las medianas y grandes empresas, este porcentaje fue de 84 %. Considerando que en la práctica sólo es posible fiscalizar a un grupo más bien reducido de empresas en un año, se recomienda realizar una combinación de los resultados obtenidos con las redes neuronales y las redes bayesianas, de manera de seleccionar para fiscalización a aquellos que aparecen catalogados como fraude en la red neuronal y que tienen las probabilidades más altas de cometer fraude según la red bayesiana.

En términos de recaudación, la predicción de un caso de fraude en una micro y pequeña empresa aporta un beneficio neto de $ 86.282, mientras que para una mediana y gran empresa, esta cifra aumenta a un $3.424.083, lo que permitiría reducir la evasión por concepto de IVA de manera significativa, si consideramos el total de casos auditados en un año.
De acuerdo a estudios que ha realizado el SII, se estima que aproximadamente un 20 % de los contribuyentes utilizan facturas para evadir impuesto. No existe información desagregada por tipo de contribuyente, pero suponiendo que este porcentaje se repite en cada segmento y considerando los porcentajes de clasificación de casos con fraude y no fraude de los modelos de red neuronal, se tiene que el universo de potenciales usuarios de facturas es de 116.000 micro y pequeñas empresas y 4.768 medianas y grandes empresas, que generan un ingreso por fiscalización de $21.344 millones de pesos y $80.102 millones de pesos respectivamente, generando un potencial de recaudación de $101.446 millones de pesos.
Finalmente, para probar la capacidad predictiva real del modelo desarrollado y siendo concordante con el punto anterior, resulta vital su aplicación en actividades que permitan determinar en terreno el nivel de acierto en la clasificación de los contribuyentes seleccionados en la muestra, para lo cual se recomienda la implementación de un programa piloto que estará dirigido a los dos segmentos económicos estudiados, que será concluyente en términos de la efectividad real del modelo.

Referencias

[1] Arnaiz, T., García, J. A. y López, J.M. Los Planes Integrales para la Prevención y Corrección del Fraude Fiscal. *Banco Interamericano de Desarrollo (BID)* . 2006.

[2] Bolton, R. y Hand, D. Statistical Fraud Detection: A Review. *Statistical Science*, Vol. 17- N°3. 2002.

[3] Centro Interamericano de Administraciones Tributarias. Métodos de Selección de Declaraciones sujetas al Control Concurrente ocupando Herramientas de Minería de Datos. *Programa Regional (TC-00-05-00-8-RG). Superintendencia Nacional de Administración Tributaria*, Perú. 2004.

[4] Clifton, P. y Chun, W. Investigative Data Mining in Fraud Detection. *School of Business Systems, Monash University.*. 2003.

[5] Davia, H.R., Coggins, J.W. y Kastantin, J. Accountant's Guide to Fraud Detection and Control (2da edición). 2000.

[6] Denny, Williams, G., Christe, P. (2007). Exploratory Multilevel Hot Spot Analysis: Australian Taxation Office Case Study. Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. *Conferences in Research and Practice in Information Technology (CRPIT)*, Vol. 70. 2007.

[7] Digimpietri, L., Trevisan, N., Meira, L., Jambeiro, J., Ferreira, C. y Kondo, A. Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System. *Proceedings of the 9th Annual International Digital Government Research Conference.* 2008.

[8] Ernst&Young *9th Global Fraud Survey 2006: Fraud Risk in emerging markets.* Junio. 2006.

[9] Fayyad, U., Piatestky-Shapiro, G., Smyth, P. From data mining to knowledge discovery in databases. *American association for artificial intelligence* 0738-4602, 37-54. 1996

[10] Government Accountability Office (GAO), United States. Data Mining: Agencies have taken key steps to protect privacy in selected efforts, but significant Compliance Issues Remain. Mayo. 2004.

[11] Government Accountability Office (GAO), United States. Lessons Learned from Other Countries on Compliance Risks, Administrative Costs, Compliance Burden and Transition. *Report to Congressional Requesters*, Abril. 2008.

[12] Harrison, G. y Krelove, R. (2005). VAT Refunds: A Review of Country Experience. *International Monetary Fund (IMF) Working Paper.* Noviembre. 2005.

[13] Luckeheide, S. Segmentación de los Contribuyentes que declaran IVA aplicando herramientas de clustering. *Revista de Ingeniería en Sistemas. Volumen XXI.* 2007.

[14] Munoz, D.J. Proceso de Reconocimiento de Objetos asistido por computador, aplicando Gases Neuronales y técnicas de Minería de Datos. *Scientia et Technica*- Año XII, No 30, Mayo. 2006.

[15] Myatt Glenn, J. Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining. *Wiley Interscience.* 2007.

[16] OECD. Compliance Measurement, Practice Note. Centre for Tax Policy and Administration, Tax Guidance Serie. *General Administrative Principles - GAP004 Compliance Measurement*- Junio. 1999.

[17] OECD. Compliance Risk Management, Use of Random Audit Programs. Forum on Tax Administration Compliance Subgroup. *Centre for Tax Policy and Administration*. Septiembre. 2004.

[18] OECD. Compliance Risk Management, Audit Case Selection Systems. Forum on Tax Administration Compliance Subgroup. *Centre for Tax Policy and Administration*. Octubre. 2004.

[19] Servicio de Impuestos Internos. Información de Cuenta Pública 2010. `htt://www.sii.cl/cuenta_publica/`. 2011.

[20] Superintendencia Nacional de Administración Tributaria. La Gestión de la Sunat en los últimos cinco años: Principales Avances y Desafíos. 2006.

[21] Tanzyi, V. y Shome, P. (1993). Tax Evasion: Causes, Estimation Methods, and Penalties a Focus on Latin America. *Documento elaborado para el Proyecto Regional de Política Fiscal CEPAL/PNUD*. 1993.

[22] Velasco, D. Redes Bayesianas. *Inteligencia Artificial II*. 2007

[23] Velázquez, J. y Palade, V. Adaptative Web Sites: A Knowledge Extraction from Web Data Approach". *Frontiers in Artificial Intelligence and Applications*, Volumen 170. 2008.

Programas de Postgrado y Postítulos DII

DOCTORADO

La carencia de capital humano avanzado es una preocupación nacional. Por ello la Universidad de Chile ha creado el Doctorado en Sistemas de Ingeniería (DSI), el cual forma profesionales de excelencia en esta área. Su foco es el desarrollo de habilidades de resolución de problemas con técnicas avanzadas y metodologías multidisciplinarias.

El programa está dirigido a estudiantes que hayan finalizado sus estudios en Economía, Ingeniería, Matemáticas u otras disciplinas con un fuerte componente cuantitativo. Los egresados contarán con herramientas y habilidades para abordar problemas complejos, que combinan gran tamaño, aleatoriedad y aspectos dinámicos y/o externalidades importantes, como consecuencia de la interacción sistémica entre agentes y procesos. Serán capaces de formular soluciones con una perspectiva que integre el conocimiento técnico con metodologías en el estado del arte, apoyando al desarrollo productivo y social de las organizaciones o permientiendo la formación de nuevos enfoques y teorías en el ámbito científico.

El DSI es impartido por los Departamentos de Ingeniería Industrial, Ingeniería Matemática, Ingeniería Eléctrica y la División de Transporte de Ingeniería Civil de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile

Más información en http://www.sistemasdeingenieria.cl/doctorado/

MAGÍSTERES

Magíster en Gestión de Operaciones MGO

El Magíster en Gestión de Operaciones (MGO) busca formar profesionales de excelencia en esquemas de gestión, uso de modelos y tecnologías de información, con capacidad de resolver problemas complejos en gestión de operaciones.

El Magíster en Gestión de Operaciones es impartido por el Centro de Gestión de Operaciones (CGO) del Departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, que recoge la vasta experiencia de sus integrantes en las áreas forestal, minera, de servicios y manufactura, tanto a través de proyectos de investigación como de consultoría.

Los egresados del Magíster en Gestión de Operaciones se desempeñan en cargos de primer nivel en empresas de servicios y manufactura nacionales e internacionales. También trabajan en universidades como académicos e investigadores, y algunos siguen doctorados en prestigiosas universidades como MIT y Columbia.

Requisitos de Admisión

Poseer el grado de licenciado en Ciencias de la Ingeniería o su equivalente.

Calendario de Postulaciones

Semestre Otoño
Periodo de postulaciones: de octubre a 15 de diciembre.
Inicio de clases: marzo de cada año.

Semestre Primavera
Período de postulaciones: de abril a 15 de junio.
Inicio de clases: julio de cada año.

Mayor información: julie@dii.uchile.cl | www.dii.uchile.cl/mgo | Tel: (56 2) 978 4017 | (56 2) 978 4073

INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

MBE
Master in Business Engineering

Magíster en Ingeniería de Negocios con Tecnologías de la Información

Los líderes de hoy comprenden cómo la tecnología lleva a las empresas al éxito.

A Quién está Dirigido

Ejecutivos y profesionales que deseen liderar o ejecutar proyectos innovadores de diseño integral y sistémico de los negocios orientados a mejorar su competitividad.

Metodología

Este es un Magíster integrador, conformado por un conjunto de cursos de gestión, modelos analíticos aplicados, diseño de negocios, arquitectura y procesos, tecnologías de información de base y diseño de aplicaciones, y de inducción de habilidades de innovación.

Además de las evaluaciones tradicionales por medio de controles y exámenes, una parte fundamental del trabajo de los alumnos será el desarrollo de un proyecto de innovación en el negocio de la empresa auspiciadora -donde ejecutará su residencia-, el cual se llevará a cabo durante todo el programa, en los cursos obligatorios del mismo.

Duración:
3 semestres académicos más un semestre para dar término al Proyecto de Grado.

Horario:
Martes o jueves vespertino, viernes de 14:30 a 18:00 horas y sábados de 8:30 a 11:45 horas.

Informaciones:
Coordinadora: Ana María Valenzuela.
(56 2) 978 4835 / anamaria@dii.uchile.cl

www.dii.uchile.cl

fcfm FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

Magíster en Economía Aplicada MAGCEA

El Magíster en Economía Aplicada (MAGCEA) busca formar profesionales de gran competencia analítica y una sólida base en economía.

Es impartido por el Centro de Economía Aplicada (CEA) del Departamento de Ingeniería Industrial, de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. El CEA es líder en investigación en economía en Chile y en el desarrollo de propuestas para las políticas públicas.

Muchos graduados del programa han seguido estudios de doctorado en universidades como Harvard, Stanford, MIT, Princeton y Yale, entre otras.

Nuestros egresados son altamente requeridos en el mercado laboral, se desempeñan en empresas líderes a nivel nacional e internacional, en destacados organismos estatales y en importantes organismos internacionales.

Requisitos de Admisión

Poseer un título profesional, nacional o extranjero, que exija al menos cinco años de estudios, o el grado de licenciado en campos disciplinarios afines a la especialidad.

Calendario de Postulaciones

Semestre Otoño
Período de postulaciones: de octubre a 15 de diciembre.
Inicio de clases: marzo de cada año.

Semestre Primavera
Período de postulaciones: de abril a 15 de junio.
Inicio de clases: julio de cada año.

Mayor información: magcea@dii.uchile.cl | www.magcea-uchile.cl | Tel: (56 2) 978 4084 | (56 2) 978 4073

Magíster en Gestión y Políticas Públicas MGPP

El Magíster en Gestión y Políticas Públicas, tiene como propósito la formación avanzada de profesionales interesados en la formulación y ejecución de políticas públicas.

El MGPP forma líderes y servidores públicos del más alto nivel, capaces de conceptualizar, pensar y discutir sus visiones e ideas sobre el futuro de América Latina.

Se imparte en dos horarios: Diurno y Ejecutivo.

Características Distintivas

* Excelencia Académica
* Cuerpo docente de primer nivel
* Orientado a profesionales de formación diversa
* Alta tasa de graduación (80%)
* Reconocido entre los mejores en su área en América Latina
* Acreditado por el CNA
* 17 años formando líderes

Requisitos de Admisión

* Poseer el grado de licenciado o título universitario

Horario Diurno:
Inicio: junio de cada año
Duración: 19 meses

Horario Ejecutivo:
Inicio: julio de cada año
Duración: 24 meses

Postulaciones:
Hasta el **15 de noviembre** para personas que postulan a becas de instituciones

Hasta el **15 de abril** para personas que cuentan con fondos propios

Mayor información: mgpp@dii.uchile.cl | www.mgpp.cl | Tel.: (562) 978 4067 | Fax 689 4987

Magíster en Gestión para la Globalización

El Magíster en Gestión para la Globalización es el único MBA realmente global de Chile. Tiene una duración de 18 meses, de los cuales 8 se cursan en una de 6 escuelas de negocios en convenio entre EE.UU., Inglaterra y Australia. Posteriormente, los alumnos completan su estadía en el extranjero con un study tour por Asia Pacífico.

Impartido en alianza entre el Departamento de Ingeniería Industrial de la Universidad de Chile y Minera Escondida (operada por BHP Billiton), el programa apunta a preparar capital humano de clase mundial para enfrentar los desafíos y aprovechar las oportunidades de la globalización, y de esta manera contribuir al desarrollo del país.

Está dirigido a jóvenes profesionales que aspiran convertirse en los futuros líderes con visión global de negocios de Chile

Los interesados en postular a la sexta versión de este exitoso Magíster, deben participar en un riguroso proceso de selección centrado en la excelencia académica y profesional, la meritocracia y la representación de la diversidad. Todos los aceptados acceden a una beca de monto variable y ayuda financiera.

Más Información
wwww.magisterglobalizacion.cl

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

MINERA ESCONDIDA
Operada por BHP Billiton


INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
Porque no da lo mismo donde estudiar
MBA
Magíster en Gestión y Dirección de Empresas
Primeros en Latinoámerica
Fortaleza Académica
Experiencia en negocios
Producción y difusión del conocimiento
Primeros en Chile
Economía, Operaciones,
Marketing, Finanzas,
Recursos Humanos
Ranking Escuelas de Negocios 2011, Revista América Economía
Programa para profesionales en general
Postulaciones Abiertas
Inicio de clases Abril 2012
(56-2) 978 4002 | mba@uchile.cl
www.mbauchile.cl
fcfm
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE


INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
MBA
Magíster en Gestión y
Dirección de Empresas
VERSIÓN INDUSTRIA MINERA
Formato Week end
Segunda Versión: Primer semestre 2012
Información: (56 2) 978 4002 | mbamin@dii.uchile.cl | www.mbamin.cl
fcfm
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

EDUCACIÓN EJECUTIVA

Creación de Material Académico

El Programa de Desarrollo de Casos del Instituto Sistemas Complejos de Ingeniería busca generar material docente innovador y herramientas que ayuden a mejorar la enseñanza de la ingeniería en Chile.

Los Casos de Estudio son documentos pedagógicos que presentan problemas basados en desafíos reales que han sido abordados por académicos del Instituto.

La aplicación de un Caso de Estudio en clases se basa en la idea de discutir el problema, identificar sus distintas aristas y posibles estrategias de solución. A diferencia de los problemas clásicos, los Casos tienen varias posibles soluciones y motivan pensamiento creativo en el alumno.

En nuestro sitio web usted podrá encontrar todos los Casos de Estudio generados hasta hoy. Las temas abordados por nuestros casos van desde el diseño de un sistema de transporte público para una ciudad hasta la programación del torneo de apertura de fútbol, abarcando áreas de ivestigación tales como combinatoria aplicada, teoría de grafos, optimización, gestión de operaciones y minería de datos entre otros.

Visite nuestro sitio web

www.isci.cl sección "Casos"
casos@sistemasdeingenieria.cl