Segmentación de los Contribuyentes que Declaran IVA Aplicando Herramientas de Clustering

Sandra Lückeheide C.*
Juan D. Velásquez*
Lorena Cerda**

Resumen

En este trabajo se llevó a cabo una caracterización de contribuyentes que declaran IVA a través de la aplicación de algoritmos de clustering, con el fin de aportar nueva información de apoyo a la labor fiscalizadora del SII. La segmentación se realizó a partir de la información tributaria consignada por los contribuyentes, en sus declaraciones de IVA (Formulario F29) y en su Inicio de Actividades.

En primer lugar, aplicando un proceso de limpieza y reducción de los datos, se confeccionó un vector de características compuesto por la información del formulario F29, del documento de Inicio de Actividades, tales como el conjunto de Actividades Económicas (actecos) y la comuna. Sobre este vector de características, se aplicaron los algoritmos de clustering Self Organizing Feature Map (SOFM) y K-means. Comparando los resultados de distintas aplicaciones de estos dos métodos, se obtuvo el vector de características final y la segmentación de los contribuyentes. En ambos casos, SOFM y K-means, los resultados de la segmentación son comparables, lo cual valida el modelo de comportamiento del contribuyente desarrollado.

A partir de la segmentación propuesta, el organismo fiscalizador mejora su labor, haciendo más certero el proceso de validación de la información que declara el contribuyente.

Palabras Clave: DATA MINING, SEGMENTACIÓN, CLUSTERING, CLASIFICACIÓN.

*Departamento de Ingeniería Industrial, Universidad de Chile
**Servicio de Impuestos Internos de Chile

1. Introducción

El Servicio de Impuestos Internos (SII) es una institución del Estado, cuya labor es la administración tributaria, siendo el principal ente fiscalizador tributario. El SII es responsable de administrar el sistema de tributos internos, facilitar y fiscalizar el cumplimiento tributario, propiciar la reducción de costos de cumplimiento y potenciar la modernización del Estado; lo anterior en pos de fortalecer el nivel de cumplimiento tributario y del desarrollo económico de Chile y de su gente.

De acuerdo a la Ley, las funciones del SII son la "aplicación y fiscalización de todos los impuestos internos actualmente establecidos o que se establecieren, fiscales o de otro carácter en que tenga interés el Fisco y cuyo control no esté especialmente encomendado por la ley a una autoridad diferente".

El SII usa actualmente diversos métodos de fiscalización. Algunos de ellos se basan en lo que el contribuyente declara en algún determinado código de un formulario, lo que conlleva a que algunos contribuyentes puedan ser fiscalizados más de una vez, al ser seleccionados en distintos tipos de fiscalizaciones. Esto resulta ineficiente para el SII y molesto para el contribuyente. Otro método se basa en lo que se esperaría que el contribuyente declare, según sus actividades económicas declaradas en su Inicio de Actividades. Este último, se determina en base a la experiencia de los fiscalizadores, es decir, de forma cualitativa y subjetiva.

Una forma alternativa factible de enfrentar este problema, es la realización de una caracterización de los contribuyentes que declaran IVA, a partir de un agrupamiento basado en su información tributaria, contenida en los formularios de Declaración Mensual y Pago Simultáneo de Impuestos (IVA, Formulario 29) y en su Inicio de Actividades. De esta forma, se buscan los mejores grupos de equivalencia de comportamiento económico, determinado por la información tributaria declarada por los contribuyentes en el formulario F29, e información cualitativa declarada en el Inicio de Actividades, tales como el conjunto de Actividades Económicas (actecos).

La segmentación de contribuyentes, constituye un apoyo a la labor fiscalizadora del SII, a través del cual se podrá analizar el comportamiento de contribuyentes con características similares, y no a la gran masa, compuesta por personas con comportamientos y actividades muy diversas. De esta forma, se puede identificar las características principales que definen a cada grupo, para luego jerarquizarlos y priorizarlos, para una fiscalización más eficiente.

2. Trabajos Previos

El rápido avance de la tecnología, la gran cantidad de datos actualmente disponible y el bajo costo relativo su almacenamiento, han incentivado la creación y el uso de distintas técnicas y algoritmos que permiten procesar los datos, extrayendo conocimientos y patrones ocultos, que de otra forma no se podrían obtener. Entre estas técnicas se encuentran las herramientas de clustering, que permiten agrupar objetos similares (y separar los objetos disímiles). Estas herramientas son actualmente muy útiles en la investigación del comportamiento humano, para el apoyo de áreas como el marketing, o en la toma de decisiones importantes.

2.1. Clustering

Las herramientas de clustering son muy populares en la extracción de patrones de conjuntos de datos, particularmente en el análisis de comportamiento humano. Esto, debido a que la formación de grupos de personas con características comunes es una tendencia natural: comunidades sociales (por ejemplo civilizaciones, países, cuyas características comunes son el idioma, raza, aspectos culturales), y dentro de estas, se forman subgrupos, por ejemplo basados en antecedentes socio-económicos. Específicamente, el análisis de cluster es muy útil en marketing, dado que las compañías buscan crear el producto preciso para un grupo específico de consumidores [17].

El análisis de clusters o clustering, también llamado segmentación de data, tiene una variedad de objetivos, todos ellos relacionados con agrupar o segmentar una colección de objetos en subconjuntos o "clusters", tal que aquellos objetos dentro de cada cluster están más cercanamente relacionados que los asignados a clusters diferentes [7]. Un objeto puede ser descrito por un conjunto de medidas, o por su relación con otros objetos. Adicionalmente, el objetivo puede ser ordenar los clusters en una jerarquía natural. Esto involucra agrupar los clusters sucesivamente, de modo que en cada nivel de la jerarquía, los clusters en un mismo grupo son más similares entre ellos, que aquellos en diferentes grupos.

Centro de todos los objetivos del clustering, es la noción de grado de similitud (o diferencia) entre los objetos individuales a ser clusterizados, y por ello es fundamental para todas las técnicas de clustering, la elección de la medida de distancia o similitud entre dos objetos. Un método de clustering intenta agrupar los objetos basados en la definición de similitud que se le provee. La situación es algo parecida a la especificación de una función de pérdida o costo, en problemas de predicción (aprendizaje supervisado). El costo asociado con

una predicción inexacta depende de consideraciones externas a la data.

Sea Ω un conjunto de m vectores $\omega_i \in \Re^n$, con $i = 1, \ldots, m$. El objetivo es particionar Ω en K grupos, donde C_j es el j-ésimo cluster. Luego, $\omega_i \in C_j$ significa que ω_i es más parecido a los elementos dentro del cluster C_j, con $j = 1, \ldots, K$ que a los elementos pertenecientes a cualquier otro cluster [17].

El clustering requiere una medida de similitud, $\zeta(w_p, w_q)$ para comparar dos vectores de Ω. La forma de determinar el número de clusters, depende del método usado.

2.1.1. Vector de Características

El vector de características (feature vector) es el conjunto de atributos (o variables) seleccionados para representar a cada objeto del conjunto de datos, luego de haberlo preprocesado, limpiado y transformado, y sobre el cual se aplican los algoritmos de Data Mining.

La elección de este vector para la aplicación de técnicas de clustering, influye directamente en los resultados del análisis, por ello es un aspecto muy importante en Data Mining, pues los resultados dependen en gran medida, de las variables consideradas en el estudio [22].

Para la aplicación de toda herramienta de data mining, se requiere generar un vector de características, compuesto por un conjunto de variables que representan las características intrínsecas del fenómeno en estudio y que luego es usado como entrada para del algoritmo de clustering. Para ello, en una primera etapa, se debe realizar una selección, limpieza y preprocesamiento del conjunto de datos y variables sobre las cuales se lleva a cabo el estudio. Es decir, dentro del conjunto de datos, se deben tratar los fuera de rango y/o inconsistentes, se transforman, normalizando las variables y los vectores, y cuando se requiere, se realiza una reducción de las variables transformadas. La reducción de dimensiones, consiste en la selección de atributos (o feature selection), es decir se selecciona el conjunto mínimo de atributos, tal que la distribución de probabilidad de las diferentes clases, dados los valores de esos atributos, sea lo más parecida posible a la distribución original, considerando los valores de todos los atributos. Existen diversos métodos de reducción de dimensiones, entre estos el Análisis de Componentes Principales [14]. Luego de esta etapa, se ha obtenido el vector de características final, al que se le aplicará los algoritmos de Data Mining, donde se comprueba si el vector seleccionado es el indicado. Por ello, esta parte puede ser iterativa, pues se debe experimentar hasta dar con el mejor vector de características.

Se debe tener en cuenta que, al usar distintas técnicas en cada uno de los pasos mencionados, se puede llegar a distintos resultados, por lo que se debe hacer una cuidadosa elección de cada técnica a aplicar.

2.2. Self Organizing Feature Maps

Self-Organizing Maps (SOFM) es uno de los modelos más populares de Redes Neuronales. Elabora una cuantización del espacio formado por los datos de entrenamiento, y simultáneamente lleva a cabo una proyección con preservación topológica en una grilla regular de baja dimensión [18].

Una red de Kohonen, o SOFM (Self-Organizing Map) es una RNA no supervisada, competitiva, distribuida de forma regular en una grilla de, normalmente, dos dimensiones, cuyo fin es descubrir la estructura subyacente de los datos introducidos en ella. A lo largo del entrenamiento de la red, los vectores de características son introducidos en cada neurona y se comparan con el vector de peso característico de cada neurona. La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora (o BMU) y ella y sus vecinas verán modificados sus vectores de pesos.

En las SOFM de dos dimensiones, se pueden distinguir dos tipos de rejillas:

- Rejilla hexagonal: en ella cada neurona tiene seis vecinos (excepto los extremos).
- Rejilla rectangular: cada neurona tiene cuatro vecinos.

Cada neurona de la red tiene asociado un vector de pesos (o prototipo) de la misma dimensión que los datos de entrada. Éste sería el espacio de entrada de la red, mientras que el espacio de salida sería la posición en el mapa de cada neurona.

Las neuronas mantienen con sus vecinas relaciones de vecindad, las cuales son claves para conformar el mapa durante la etapa de entrenamiento.

En cada paso se introduce un vector de datos en cada neurona y se calcula la "similitud" entre éste y el vector de peso de cada neurona.

$$\|X_j - m_{BMU}\| = \min_j\{\|X_i - m_j\|\} \tag{1}$$

$$m_k \leftarrow m_k + \alpha(X_i - M_k) \tag{2}$$

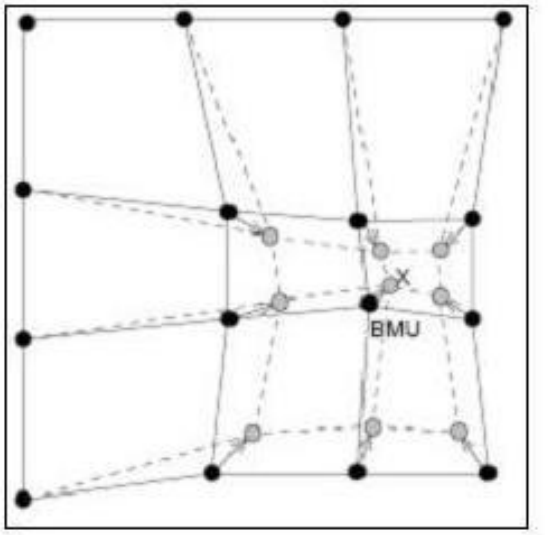
La neurona más parecida al vector de entrada, es la ganadora (o BMU, Best-Matching Unit, Unidad con mejor ajuste). Para medir la similaridad se utiliza usualmente la distancia euclideana. Tras ello, los vectores de pesos de la BMU y de sus vecinos son actualizados, de tal forma que se acercan al vector de entrada.

SOFM tiene propiedades tanto de algoritmos de cuantización de vectores, como de proyección de vectores, lo que permite hacer una reducción del conjunto de datos original, manteniendo la representatividad, además de hacer análisis posteriores como clustering o visualización.

Además de los beneficios computacionales ofrecidos por la cuantización de vectores, las principales ventajas del SOFM son [18]: a) su robustez, dado que todos los prototipos son afectados por todos los datos, b) su sintonización local, pues se trabaja en la vecindad de cada unidad del mapa, se sintoniza localmente con la densidad de los datos y c) su facilidad de visualización.

Entre sus desventajas, se destacan: los efectos de borde, dado que las definición de las vecindades no es simétrica en los bordes del mapa, y la contracción del rango de valores de las variables, en que se dejan algunos valores afuera que bajo algún punto de vista podrían ser interesantes.

Figura 1: Actualización del BMU y sus vecinos, hacia X



Fuente: [18]

2.2.1. K-Means

El algoritmo K-means, es uno de los métodos de clustering iterativos más usados. Es destinado a situaciones en las cuales todas las variables son de tipo cuantitativo, y la distancia euclideana es generalmente escogida como medida de disimilitud.

La dispersión intra-puntos puede escribirse como:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{c(i)=K}\sum_{C(i')=K}\|X_i - X_{i'}\|^2 = \sum_{K=1}\sum_{C(i)=K}\|X_i - \bar{X}_k\|^2 \quad (3)$$

Donde $\bar{X}_k = (\bar{X}_{1k}, ..., \bar{X}_{pk})$ es el vector promedio sociado al cluster K.

Luego, este criterio es minimizado asignando las N observaciones de los K clusters de tal forma que dentro de cada cluster, la disimilitud promedio de

las observaciones desde la media de los clusters definida por los puntos en ese cluster, es minimizada.

Algoritmo K-means:

1. Para una asignación de clusters dada C, la varianza de cluster total

$$\bar{X}_s = arg \min_m \sum_{i \epsilon S} \|X_i - m\| \tag{4}$$

es minimizada respecto a $m_1, ..., m_k$ dando las medias de los clusters asignados actualmente

$$C^* = \min_c \sum_{K=1}^{K} \sum_{C(i)=K} \|X_i - \bar{X}_k\|^2 \tag{5}$$

2. Dado un conjunto actual de medias m1,...,mk, (4) es minimizado asignando cada observación al cluster cuya media es la más cercana. Esto es,

$$C(i) = arg \min_{1 \leq k \leq K} \|X_i - m_k\|^2 \tag{6}$$

3. Los pasos 1 y 2 son iterados hasta que las asignaciones no cambian

Las principales ventajas de los algoritmos K-means son su simplicidad, sencillez, no es sensible al orden de los datos, y es basado en el sólido fundamento del análisis de varianzas [7]. Entre las desventajas, se cuentan la fuerte dependencia del resultado de la asignación inicial de los centroides, el óptimo encontrado es local y puede estar bastante lejos del global, la dificultad de una buena elección en el número de clusters a encontrar, es un proceso sensible a los datos fuera de rango, el algoritmo carece de escalabilidad, se limita a abarcar sólo atributos numéricos y los clusters resultantes pueden ser desequilibrados.

3. Segmentación de Contribuyentes

A continuación se describe la elaboración del vector de características, la aplicación de dos herramientas de clustering, Self Organizing Feature Maps y K-means, y la comparación de los resultados de ambos métodos.

La herramienta utilizada para usar los algoritmos es R[1] , un paquete Open Source estadístico y de Data Mining.

[1] www.r-project.org

3.1. Construyendo el vector de características

Inicialmente, los datos usados para la realización de este estudio, correspondieron a la información presentada en el año 2005, por los contribuyentes que declaran IVA (Impuesto al Valor Agregado), en el formulario F29 (Declaración Mensual y Pago Simultáneo de Impuestos), y en el formulario de Inicio de Actividad Económica. El número de contribuyentes considerados en un principio es de 597.082, y se tomaron en cuenta gran parte de códigos del formulario F29.

La declaración de IVA se hace mensualmente. Por lo tanto, para transformar los datos de mensual a anual, y comenzar a reducir de esta forma la dimensionalidad del vector de características, se decidió considerar estadísticos por cada código, como el número de no nulos en el año, la suma de los montos mensuales, el promedio y la desviación estándar (considerando sólo los meses declarados), para cada una de las variables seleccionadas.

Luego de consolidar la información, se hizo una selección y preprocesamiento de los datos. En esta etapa, se realizó la limpieza, eliminando los datos fuera de rango (outliers) y aquellos considerados inconsistentes, excluyendo de esta forma aproximadamente el 6 % de los registros iniciales. Después de la limpieza, se llevó a cabo la reducción de los datos y selección de las variables.

Dada la gran dimensionalidad del problema, tanto en número de registros (contribuyentes) como en cantidad de dimensiones, se hizo indispensable la reducción de éstas, para posibilitar y hacer más eficiente el análisis. Debido a que muchos de los códigos del fromulario de declaración de IVA, en un gran porcentaje de registros, se encuentran en blanco, proporcionando así muy poca información relevante en la discriminación de grupos, se optó por considerar sólo los códigos en los que al menos un 10 % de los contribuyentes tienen valores no nulos. Por lo tanto, teniendo la base de datos limpia, a estos códigos (o más bien a los estadísticos de estos códigos) se les aplicó un Análisis de Componentes Principales (ACP), y se procedió a seleccionar las variables (códigos del formulario) que mejor representen la variabilidad de los datos.

Considerando que las variables pueden tener distintas escalas, que puede conllevar a que aquellas con un mayor rango de valores le quiten importancia a otras con un menor rango, todas las variables consideradas fueron escaladas según la normalización "Min-Max", en el rango [0,1], según la fórmula y'= ((y-min) / (max - min)) (max'- min') + min'. Además, se tomó en cuenta que al normalizar las columnas, se pierde la relación original entre los componentes de cada vector o fila. Por ello, se llevó a cabo la normalización de los vectores (norma 1), es decir, se calculó el módulo de cada vector, y cada una de sus componentes se dividió por este valor. Para ello fue necesario extraer aquellos contribuyentes que solo tenían valores nulos en todas las variables seleccionadas (es decir, vectores de norma 0), que representan alrededor del

12 %.

La medida de distancia seleccionada para la aplicación de los algoritmos de clustering en el vector de características generado, fue la Euclidiana, que por ser la más comúnmente utilizada, viene por defecto en la mayoría de los algoritmos en R (como en gran parte de las herramientas de data mining).

Para la incorporación de la actividad económica de los contribuyentes y su comuna, se decidió cuantizar esta información, en base a lo que cada actividad y cada comuna, en promedio, genera en impuestos.

1. Primer Experimento:
   Se usaron las 10 primeras componentes principales.

   Se extrajo una muestra aleatoria de 200 mil contribuyentes, y sus respectivos valores en cada una de las variables del vector de características (las 10 componentes principales que explican el mayor porcentaje de la varianza). Se quitó de esta muestra aquellos contribuyentes cuyo vector tuviese norma 0 (es decir todas las variables con valor nulo), que corresponden a 24.599 datos, por lo tanto la muestra empleada en el análisis tiene un tamaño de 175.401 contribuyentes, que tienen al menos una variable con valor positivo.

   Luego de normalizar cada variable, según la normalización Min-Max y la normalización de los vectores, se aplicó el algoritmo K-means, con 8 clusters como condición inicial, y 10 semillas iniciales.

   Como resultado, se obtuvo 8 clusters, con los siguientes tamaños:

Cuadro 1: 8 clusters y sus tamaños

Cluster	Tamaño		Cluster	Tamaño
1	61 (0,03 %)		5	133 (0,076 %)
2	170.848 (97,4 %)		6	48(0,027 %)
3	2.667 (1,52 %)		7	111 (0,063 %)
4	563(0,32 %)		8	970 (0,553 %)

Fuente: Elaboración propia

   Se puede observar que mediante este método, se creó un gran cluster, que abarca más del 97 % de la muestra, y los clusters restantes contienen el otro 3 % (con diferencias considerables de tamaño entre algunos de ellos también).

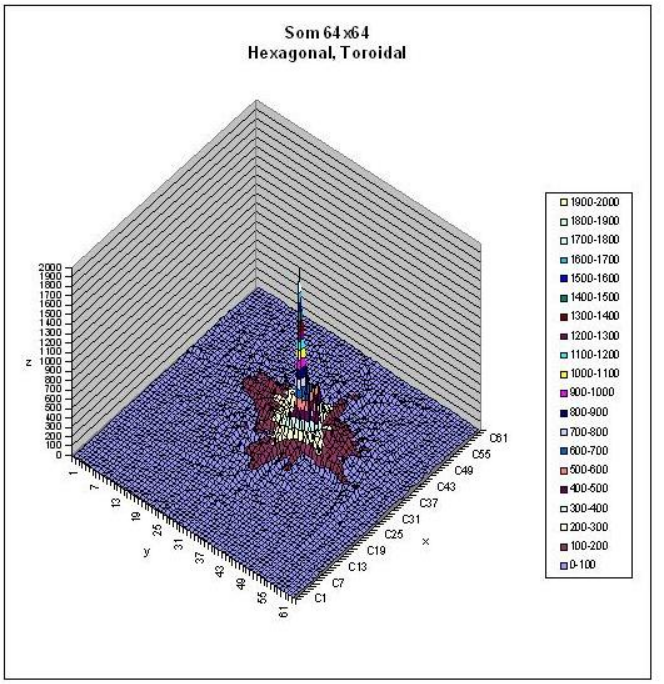
   Se probó con distintos números de clusters, entre 3 y 20 clusters, obteniéndose los mismos resultados anteriores: un solo grupo contiene más del 90 % de los datos.

Para verificar si el problema era el método empleado, se probó aplicando el algoritmo SOFM, al mismo conjunto de datos.

Utilizando el paquete "som" de R , se aplica el algoritmo SOFM a la misma muestra anterior, de 175.401 contribuyentes, y las primeras 10 Componentes Principales. Como se observa en la Figura 2, mediante este análisis se obtuvo, al igual que con el Kmeans, un gran cluster que concentra la mayoría de los datos.

Considerando la posibilidad de que se estén incluyendo observaciones que produzcan ruido (que no aportan información, sino sólo distorsionan los resultados) en el estudio, se tomaron sólo los datos pertenecientes al "gran" cluster, y sobre esta nueva muestra (de un tamaño de 148.681 observaciones), se aplicaron nuevamente el método Kmeans, para evaluar si existe algún cambio en los resultados.

Figura 2: Mapa SOFM 64x64, topología hexagonal y cerrada (toroidal), muestra de 175.401 contribuyentes, 10 componentes principales (eje z y colores corresponden a número de observaciones por celda).



Fuente: Elaboración propia

Se aplicó el algoritmo Kmeans a esta nueva muestra de 148.681 observaciones, con 8 clusters como condición inicial, y 15 semillas iniciales.

Como resultado, se obtuvo 8 clusters, con los tamaños que se muestran en el Cuadro 2.

Cuadro 2: 8 clusters y sus tamaños (Análisis incluyendo sólo observaciones pertenecientes al cluster gigante)

Cluster	Tamaño
1	258
2	1.783
3	305
4	1.028
5	135.315
6	189
7	6.567
8	491

Fuente: Elaboración propia

En el Cuadro 2 se puede ver que el resultado fue similar a los anteriores: el cluster 5 contiene 135.315 observaciones, que corresponde a más del 90 % de la muestra. Se probó con distintos números de clusters, obteniendo resultados similares.

En el mundo real, los contribuyentes no se comportan igual en absoluto. A pesar de ello, los resultados de los experimentos anteriores, al entregar un gran cluster que abarca la mayoría de la muestra, muestran lo contrario. Esto podía deberse a una mala elección del vector de características, que no definía comportamientos distintos entre contribuyentes. Por ello, se experimentó luego con un nuevo vector de características. En esta ocasión, en vez de usar las componentes principales como variables, se usaron las variables originales que tengan mayor peso en las componentes principales que explicaban mayor varianza.

2. Segundo Experimento:

Se utilizaron como vector de características, los mismos 14 códigos del formulario 29, usados para el análisis de componentes principales (aquellos que tienen un porcentaje de contribuyentes con valor positivo mayor a 10 %), además del número de declaraciones del formulario en el año 2005, pero en este casom las variables originales que tenían mayor importancia en las CP. Para cada código y contribuyente, se toma en cuenta la suma total del año 2005 (en pesos) y el número de meses en que el monto es mayor que cero.

Se realizó un análisis de componentes principales, con un total de 29 variables (la cantidad de declaraciones en el año y, para los 14 códigos, suma y número de no nulos).

La importancia por código resultó bastante similar al primer análisis realizado. Pero al seleccionar las variables puras en vez de las componentes principales, se puso atención a las correlaciones entre las variables, para no trabajar con variables muy correlacionadas que distorsionen el análisis.

Luego de analizar las correlaciones, las 16 las variables seleccionadas para el análisis fueron: c142s, c142nn, c111s, c538s, c538nn, c511s, c511nn, c525s, c525nn, c504s, c504nn, c48s, c48nn, c151s, c151nn y la cantidad de declaraciones.

Del conjunto de datos inicial, se consideraron aquellos datos en que al menos una variable era no nula, que correspondía a una muestra de 173.935 contribuyentes. Se normalizó las variables mediante la normalización "Min-max", y luego se normalizó los vectores, a módulo unitario.

Se aplicó el algoritmo Kmeans, imponiendo 8 clusters como condición inicial. Los tamaños de los grupos generados por este método, se observan a continuación:

Cuadro 3: Tamaño 8 clusters (16 Variables)

Cluster	Tamaño
1	21.905
2	16.689
3	13.745
4	11.429
5	9.667
6	50.972
7	30.360
8	19.168

Fuente: Elaboración propia

Tan sólo observando el tamaño de los clusters, se nota un cambio drástico respecto a los experimentos anteriores, en los que se consideraba como variables las componentes principales. En esta prueba, el cluster de menor tamaño contiene casi el 5.5 % de la muestra, y el de mayor tamaño el 29 %.

Al observar el Cuadro 4 correspondiente a los vectores de los centros de los clusters, llama la atención la gran importancia de las variables correspondientes al número de no nulos de cada código, no así del monto total declarado en los mismos.

Cuadro 4: Centros de los 8 clusters (16 variables)

Cluster	c142nn	c142s	c111s	c538nn	c538s	c511nn	c511s	c525nn
1	0.007	0.000	0.000	0.075	0.000	0.021	0.000	0.003
2	0.012	0.000	0.000	0.488	0.001	0.059	0.000	0.011
3	0.013	0.000	0.000	0.026	0.000	0.005	0.000	0.001
4	0.069	0.001	0.004	0.345	0.013	0.221	0.001	0.075
5	0.546	0.003	0.001	0.118	0.000	0.011	0.000	0.003
6	0.009	0.000	0.001	0.546	0.001	0.024	0.000	0.007
7	0.007	0.000	0.000	0.671	0.001	0.031	0.000	0.006
8	0.070	0.000	0.002	0.512	0.002	0.421	0.000	0.026
	c525s	c504nn	c504s	c151nn	c151s	c48nn	c48s	cant
1	0.000	0.661	0.002	0.023	0.000	0.004	0.000	0.673
2	0.000	0.477	0.002	0.099	0.001	0.005	0.000	0.632
3	0.000	0.077	0.000	0.560	0.006	0.008	0.000	0.668
4	0.002	0.104	0.005	0.388	0.024	0.466	0.011	0.482
5	0.000	0.032	0.000	0.330	0.009	0.074	0.001	0.583
6	0.000	0.065	0.000	0.546	0.001	0.003	0.000	0.562
7	0.000	0.042	0.000	0.023	0.000	0.002	0.000	0.642
8	0.000	0.073	0.000	0.399	0.002	0.008	0.000	0.518

Fuente: Elaboración propia

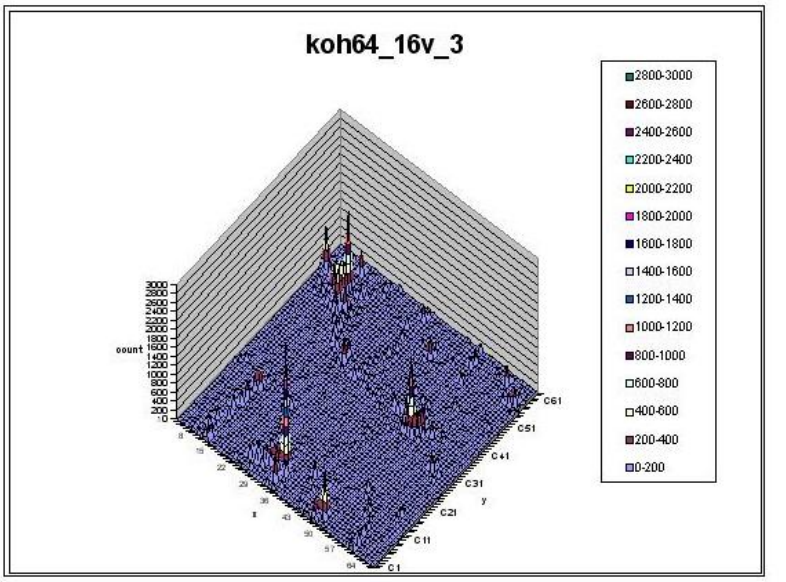
Para complementar el análisis hecho con el método Kmeans, se realizó un análisis de SOFM a la misma muestra de datos, y con las mismas variables. Utilizando el paquete "kohonen" de R, se aplica sobre esta muestra el método SOFM, con una grilla de 64x64 celdas, topología hexagonal y cerrada (toroidal), obteniéndose un mapa como el de la Figura 3.

El mapa generado por el SOFM, resultó bastante diferente a aquellos de los experimentos anteriores. En este caso, se distinguen ciertos agrupamientos visiblemente diferenciados y algunos bastante más concentrados que otras. Analizando las características de cada agrupamiento, se observó que estos se diferencian entre sí principalmente por las variables relacionadas con el número de "no nulos" (es decir cuantas veces declaró un valor positivo en el código, en el año).

De acuerdo a lo anterior, se pudo concluir que con el vector de características empleado en este experimento, ambos métodos (Kmeans y SOFM) agruparon contribuyentes basados principalmente en el número de valores no nulos para cada código, sin considerar el monto en pesos. Esto se debe a que gran parte de las observaciones tiene valores 0 o 12 en las variables "no nulos", y aquellos que tienen valores entre 1 y 11 se distribuyen uniformemente , en cambio para las variables relativas a sumas (montos), la mayoría se concentra en valores relativamente muy bajos y unos pocos en valores muy altos, por lo tanto en el análisis, luego

de la normalización de las variables y de los vectores, son los "no nulos" los que más pesan y le restan importancia a las sumas.

Figura 3: SOFM de 64x64, topología hexagonal y toroidal, muestra de datos con al menos 1 código no nulo, 16 variables (eje z y colores corresponden a número de observaciones por celda).



Fuente: Elaboración propia

Sin embargo, se consideró el hecho de que el monto declarado por el contribuyente en un determinado código, es más importante que el número de veces que declare un monto no nulo en el año . Para entender mejor el problema que puede generar la presencia de las variables "nn" en el análisis, se ejemplifica con el siguiente caso:

Suponiendo que se tiene un apicultor que produce miel (puede aplicarse a productores de cualquier otro producto) y que vende su producción anual en lotes, a grandes supermercados, los cuales realizan 4 pedidos en el año. A este contribuyente le corresponde declarar sus ventas en el código c502 (Facturas Emitidas), por lo que en la variable c502s tendrá un monto (en pesos) relativamente grande, y en la variable c502nn tendrá un valor de 4 (4 meses en que emitió facturas). Sin embargo este apicultor, todos los meses vende en su domicilio una reducida cantidad de miel a clientes de paso, a los que les entrega boleta por estas ventas. Estas ventas significan un porcentaje despreciable en comparación a lo que vende a supermercados, aún así declarará en el código c111 (Boletas), por lo que la variable c111s tendrá un valor pequeño en comparación al valor en la variable c502s y la variable c111nn tendrá un valor de 12

(dado que entregó boletas los 12 meses del año). Luego de normalizar las variables y vectores, la diferencia que existía entre el c502s y el c111s se torna despreciable y aquella entre la variable c502nn y c111nn se vuelve a su vez muy importante, y determinará la diferencia de comportamiento entre este contribuyente y los demás. Finalmente, se concluirá que este contribuyente se dedica principalmente a la venta directa, cuando en realidad ocurre todo lo contrario.

Por esta razón, se decidió sacar del análisis las variables relativas a la cantidad de "no nulos", dado el ruido que provocaban, así como también la cantidad de declaraciones en el año, dejando sólo las sumas (en pesos) para cada código, es decir 8 variables.

De esta forma, se selecciónó la suma del año 2005 de los siguientes códigos: Del recuadro Débitos y Ventas:

- c142 (Ventas y/o Servicios prestados Internos Exentos o No Gravados).
- c111 (Boletas)
- c538 (Total Débitos).

Del recuadro Créditos y Compras:

- c511 (IVA por documentos electrónicos recibidos)
- c525 (Facturas activo fijo)
- c504 (Remanente Crédito Fiscal mes anterior)

Y finalmente del recuadro Impuesto a la Renta:

- c48 (Retención Impuesto único a los Trabajadores)
- c151 (Retención de Impuesto con tasa de 10

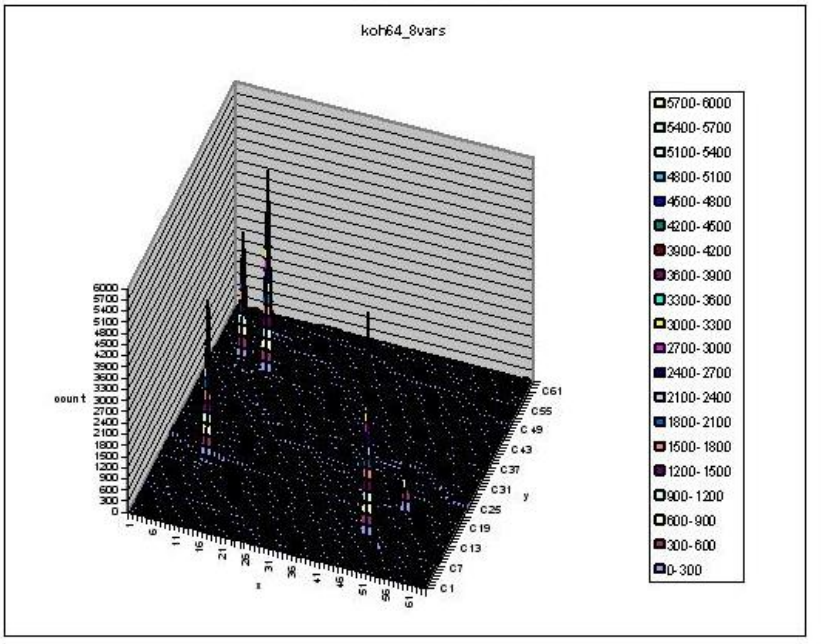
Se incluyó también la actividad económica y la comuna, pero nuevamente esto llevaba a la agrupación de la mayoría de los datos en un solo gran cluster, por lo que se decide no incorporarlas pues introdujeron más ruido que información relevante para la discriminación de grupos.

3.2. Aplicando el SOFM

Utilizando el paquete "kohonen" de R, se aplica el método SOFM, con una grilla de 64x64, de topología hexagonal y cerrada (toroidal), sobre una muestra (normalizada) de 100 mil contribuyentes.

En este caso, como se observa en el mapa generado por el SOFM, con el vector de características de 8 variables (Figura 4), se distinguen claramente 5

Figura 4: SOFM de 64x64, topología hexagonal y toroidal, 8 variables (eje z y colores corresponden a número de observaciones por celda).



Fuente: Elaboración propia

"peaks", correspondientes a celdas con una gran concentración de observaciones. En primera instancia, se consideraron estas celdas como centroides de los posibles clusters.

Para cada una de estas concentraciones, se analizaron las características tributarias de sus contribuyentes (sus declaraciones en el formulario de declaración de IVA), para determinar las similitudes existentes dentro de cada una, obteniéndose lo siguiente:

- Cluster 1: tiene una media alta en los códigos c504 (Remanente Crédito Fiscal mes anterior) y c537 (Total Créditos), y valores nulos en todos los otros códigos. En el código c91, que corresponde al total de impuesto a pagar, posee media = 0. Se puede decir que este grupo corresponde a contribuyentes que obtuvieron pérdidas en el año 2005 y se etiquetó como "Remanentes".

- Cluster 2: se caracteriza por tener montos positivos en los códigos c111 (Boletas), c538 (Total Débitos), c520 (Facturas recibidas del giro y Facturas de compra emitidas), c537 (Total Créditos), c62 (PPM 1ª Categoría), y consecuentemente en el c91 (Total a pagar). Este centroide puede estar constituido por contribuyentes que realizan actividades de venta directa, dado que declaran en el código "Boletas" (consecuentemente en

el código "Total Débitos"), y en el de "Facturas Recibidas" (consecuentemente en el código "Total Créditos"), y el código que corresponde al Pago Provisional Mensual (PPM) de 1ª Categoría. Por lo tanto, este grupo se etiquetó como "Ventas Directas".

- Cluster 3: todos los contribuyentes tienen un valor positivo en el código c142 (Ventas y/o Servicios prestados Internos Exentos, o No Gravados), y una gran parte tiene valores positivos en el c62. Luego, este grupo se llamó "Exentos".

- Cluster 4: en esta celda, los contribuyentes declaran principalmente en el código c151 (Retención de Impuesto con tasa de 10 % sobre las rentas), y en el resto de las variables poseen, en su mayoría, valores nulos. Por lo tanto, este grupo se denominó "Retenedores".

- Cluster 5: los contribuyentes de este centroide, se caracterizan por tener valores positivos en los códigos c502 (Facturas emitidas), c538 (Total Débitos), c520 (Facturas recibidas del giro y Facturas de compra emitidas), c537 (Total Créditos), c62 (PPM Neto Determinado) y c91 (Total a Pagar). Por lo tanto, está constituido por contribuyentes que realizan actividades de venta indirecta, dado que declaran en el código "Facturas Emitidas", por lo que se etiquetó como "Ventas Indirectas" o "Mayoristas".

Una vez definidos los clusters, se debió probar el clasificador. Dado que se había usado una muestra de 100 mil contribuyentes, de un total de 173.935 cuyas variables fueron inicialmente normalizadas, se seleccionaron aleatoriamente 30 mil contribuyentes del grupo que no fue considerado en el proceso de clustering, y mediante la función "map" del paquete "kohonen" de R, estas 30 mil observaciones fueron dispuestas en el mapa entrenado por las 100 mil originales.

Al colocar una nueva muestra de datos sobre el mapa entrenado inicialmente por la muestra de 100 mil datos, se generó un mapa muy similar. Las mismas celdas seleccionadas como centroides en el mapa original, en este caso también formaron "peaks" en el mapa, por su gran concentración de observaciones y se observó que los contribuyentes de cada una de estas celdas tenían características similares a los de la muestra original.

Luego de caracterizar o etiquetar cada centroide, y por ende cada cluster, se procedió a asignar todos los contribuyentes al cluster que más se le asimile, según la información contenida en él. La medida de distancia utilizada para encontrar el cluster más cercano a cada vector, fue la distancia Euclideana.

3.3. Aplicando el K-Means

Utilizando el paquete "Kmeans" de R, se aplicó el algoritmo Kmeans a la misma muestra de 100.000 contribuyentes, tomada de la muestra inicial de tamaño 173.935, con 5 clusters como condición inicial, y 20 semillas iniciales (es decir 20 pruebas con distintos centros de clusters iniciales, de las que se escoge la que entrega el mejor resultado), obteniéndose lo siguiente:

Analizando los vectores correspondientes a los centros de los clusters, se observa:

- Cluster 1: se encuentran valores altos en la variable c504s, y un poco más bajos en la variable c538s.
- Cluster 2: se encuentran valores altos en la variable c111s, y valores menores en las variable c538s y c151s.
- Cluster 3: predominan valores altos en la variable c142s y en menor medida en la variable c151s,
- Cluster 4: predomina la variable c151s y en menor medida la variable c538s.
- Cluster 5: predomina la variable c538s, y valores menores en la variable c151s.

Cuadro 5: Tamaño 5 clusters (8 Variables)

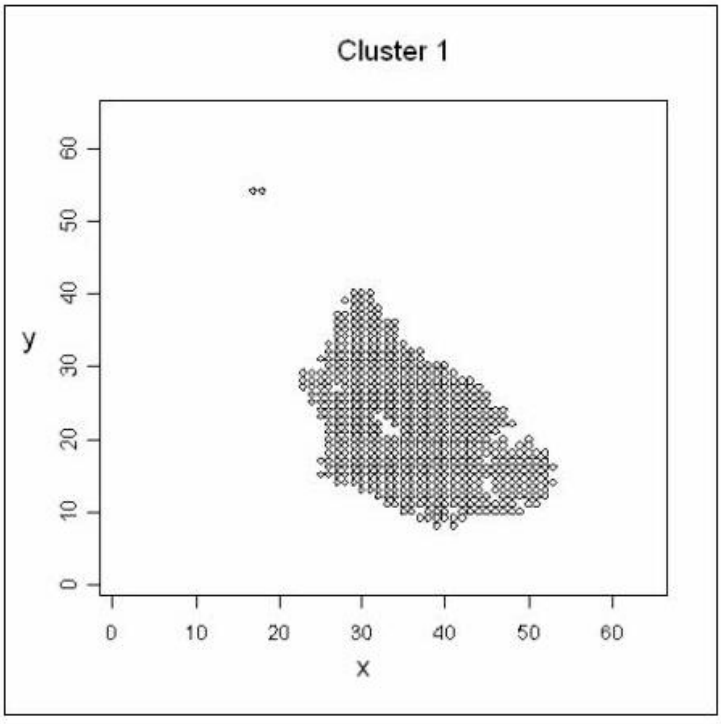
Cluster	Tamaño
1	15.583
2	32.797
3	3.405
4	27.662
5	20.533

Fuente: Elaboración propia

3.4. Comparación de Resultados

Los vectores de los centros de los clusters generados por el Kmeans, resultaron con características muy similares a aquellos de los centros del SOFM: valores relativamente altos en la variable c504s en el primero, valores altos de la variable c111s en el segundo (y consecuentemente también cierta presencia del código c538), presencia importante de la variable c142s en el tercer cluster, de la variable c151s en el cuarto y finalmente valores altos de la variable c538s en el quinto.

Luego, se confirmó gráficamente si los contribuyentes unidos por el método Kmeans, se encuentran unidos en el mapa generado por el SOFM. Para ello, se elaboraron gráficos, en los que se observa la ubicación en el mapa generado por el SOFM, de los contribuyentes de cada cluster formado por el Kmeans. A modo de ejemplo, en la Figura 5 se muestra el Cluster 1 ("Remanentes") generado por el método Kmeans, y su ubicación en el mapa del SOFM. Se puede ver que, a excepción de un par de celdas, todas las celdas que contienen contribuyentes del Cluster 1 del Kmeans, se encuentran juntas en el SOFM. Algo similar se puede concluir respecto a los demás clusters. Por lo tanto, se puede concluir que el agrupamiento estuvo bien hecho, al llegar a resultados similares, por caminos diferentes.



Fuente: Elaboración propia

4. Aplicación del Clasificador

Luego de comprobar que efectivamente existían diferencias en el total de tributación (de IVA) por grupo y de evaluar la calidad de la segmentación, se procedió a generar indicadores que permitan, dentro de cada grupo, encontrar aquellos contribuyentes cuyo comportamiento se aleje significativamente del

resto del grupo.

En general, se esperaría que para cualquier contribuyente, la suma de los montos en el código c538 (Total Débitos) sea superior a la suma de los montos en el código c537 (Total Créditos), es decir que los ingresos deben ser superiores a los egresos . En primer lugar, se creó el indicador "Débitos / Créditos". Dado que algunos tienen valor cero en Créditos, se debió transformar a "Débitos / (Créditos + 1)".

La razón "Débito/Crédito" (o Débito - Crédito <= 0) es usualmente utilizada como uno de los métodos de fiscalización. Sin embargo, este cálculo no tiene mucho sentido para ciertos tipos de contribuyentes, como los del cluster 3 ("Exentos"), que en su mayoría corresponden contribuyentes cuyas actividades no generan débito fiscal, o el cluster 1 ("Remanentes"), donde la media de la suma de Débitos es menor a la de Créditos. Pero este indicador sí tiene sentido principalmente para aquellos del cluster 2 ("Ventas Directas") y del 5 ("Ventas Indirectas"), en los que en el primer cuartil, este indicador ya tiene un valor mayor a 1 (es decir que en estos casos, la mayoría cumple que Débitos < Créditos). Por lo tanto, para los contribuyentes pertenecientes a estos clusters, se debe poner principal atención en aquellos que tienen un valor inferior a 1 (o un valor levemente mayor a 1) en el indicador.

Luego, a partir del indicador creado, se generó una medida del comportamiento de pago de cada uno de estos 2 grupos, que son aquellos que tienen más incentivo (dado que todas las transacciones que realizan se encuentran gravadas) y oportunidades de evadir. El cluster 3 ("Exentos") no es tan interesante de analizar en este sentido, pues dado que las actividades que realizan la mayor parte se sus contribuyentes se encuentran exentas, estos no tienen mayor incentivo a distorsionar los montos que declaran. El cluster 1 ("Remanentes"), puede ser sujeto a mayor estudio en trabajos posteriores, pues corresponde a contribuyentes en que los valores mas importantes corresponden al código c504 (Remanente Crédito Fiscal mes anterior), es decir en su mayoría son contribuyentes que declaran pérdidas, y por ello la mayor parte tiene valor nulo en el código c91 (Total a Pagar).

Por lo tanto, considerando sólo los clusters 2 ("Ventas Directas") y 5 ("Ventas Indirectas"), para cada uno de ellos se extrae el valor del primer cuartil en el conjunto de contribuyentes cuyo indicador es superior a uno. Estos valores se consideran como el valor mínimo esperado para la razón Débito/Crédito, que deberían tener todos los contribuyentes según el cluster al que pertenezcan.

Luego, para ambos clusters, se calculó el promedio de tributación (usando el código 91, Total a Pagar) de aquellos cuyo indicador se encontrara muy cercano a los valores calculados en la etapa anterior.

A continuación, para crear una medición del comportamiento de pago en cada cluster, para cada contribuyente cuya razón "Débito/Crédito" es menor al umbral calculado para el cluster (es decir primer cuartil), se calculó la di-

ferencia entre el valor declarado por éste en el código c91 y el promedio en el código c91 calculado en la etapa anterior, para el cluster correspondiente, que es considerado como el valor mínimo esperado a pagar. Finalmente, la suma de las diferencias entre el valor real y el esperado, calculadas para cada contribuyente de un determinado cluster, corresponde al indicador de comportamiento de pago del mismo. Mientras más alto es el valor de éste, más diferencia (negativa) hay entre lo que pagan los contribuyentes "bajos" (cuyo indicador Débito/Crédito es bajo).

De esta forma, además de generar una "alarma" en aquellos contribuyentes bajo el umbral determinado, se obtuvo la diferencia total para cada cluster considerado, entre lo que estos pagan y lo que se esperaría que paguen.

5. Conclusiones y Trabajos Futuros

A través de este trabajo, se realizó una caracterización de los contribuyentes que declaran IVA usando su información tributaria del año 2005. De esta forma, obtuvo información novedosa y potencialmente útil para el SII, en particular en el proceso de selección de contribuyentes a fiscalizar.

La elección del vector de características es fundamental, en este y en la mayoría de los trabajos de Data Mining, lo que quedó claramente demostrado, al obtener resultados absolutamente diferentes entre una y otra elección de vector de características, a veces incluso bajo pequeños cambios. Por lo tanto, la elección del vector determina en gran medida el resultado final del análisis. Luego de varios experimentos, se concluyó que el vector de características que mejor discrimina entre los contribuyentes, dada la calidad del clustering resultante, es aquel compuesto por los siguientes códigos, declarados en el Formulario de Declaración Mensual y Pago Simultáneo de Impuestos: c142 (Ventas y/o Servicios prestados Internos Exentos o No Gravados), c111 (Boletas), c538 (Total Débitos), c525 (Facturas Activo Fijo), c511 (IVA por documentos electrónicos recibidos), c504 (Remanente Crédito Fiscal mes anterior), c48 (Retención Impuesto único a los Trabajadores) y c151 (Retención de Impuesto con tasa del 10 %).

Usando el vector de características seleccionado, se agruparon los contribuyentes, utilizando los algoritmos K-means y SOFM. Se seleccionó este último, obteniéndose 5 grupos claramente diferenciados respecto a los montos declarados en diferentes códigos del formulario F29. Mediante un análisis estadístico, se verificó que estos grupos son significativamente diferentes respecto al Impuesto Total a pagar.

Para caracterizar el comportamiento de un contribuyente dentro de su grupo, se creó un indicador (razón entre Débitos y Créditos), que resulta útil

principalmente en 2 de los clusters encontrados (Cluster de "Ventas Directas" y el de "Ventas Indirectas" por tener estos grupos mayor incentivo y oportunidades de evadir impuestos. A partir de esto, se generó otro indicador, caracterizando el comportamiento de pago de los contribuyentes de cada grupo.

Quedó además demostrado que existen otras formas de agrupar a los contribuyentes, además de las que actualmente se conocen como el tamaño de la empresa o el rubro o sector al cual pertenezca. En este caso, la agrupación se hizo en base a los códigos que declaran (independiente de los montos declarados en ellos), obteniéndose un grupo caracterizado por generar pérdidas ("Retenedores"), otro grupo consistente en los que venden directamente al consumidor final ("Ventas Directas"), otro en que los contribuyentes realizan actividades exentas ("Exentos", entre los que se incluyen por ejemplo los centros médicos), otro compuesto por los contribuyentes intermediarios ("Ventas Indirectas") y el grupo donde se reúnen los contribuyentes que emplean y retienen ("Retenedores").

La metodología generada en este trabajo para agrupar contribuyentes de comportamiento similar, resulta bastante confiable y perpetuable en el tiempo. Esto, debido a que ella no es tan sensible a los montos declarados en cada código, sino más bien al código en sí mismo, es decir, si este es usado o no por el contribuyente.

A partir de estas conclusiones, se proponen las siguientes recomendaciones:

Resulta interesante realizar un análisis más profundo del cluster de "Remanentes" (cluster 1), correspondiente a aquellos contribuyentes que declaran pérdidas, dado que se comprobó que es un grupo bastante importante en cuanto a tamaño (15,7 % de los contribuyentes considerados) y que en este estudio no fue posible de caracterizar y estudiar con profundidad. Sobre todo, se debería estudiar el comportamiento de los contribuyentes pertenecientes a este grupo, en los años anteriores y posteriores, tomando como hipótesis el hecho de que hay incentivo a permanecer en una actividad, solo mientras las utilidades son positivas.

En trabajos futuros, se puede elaborar otros indicadores, que permitan evaluar el comportamiento en los otros clusters, como el 3 ("Actividades Exentas") y el 4 ("Retenedores").

Se recomienda realizar este estudio, usando otros métodos para el preprocesamiento de las variables y el agrupamiento de los datos, que eventualmente podrían llevar a resultados diferentes. Para ello se puede por ejemplo: cambiar la topología del SOFM en su forma y tamaño de la grilla, re-muestrear, usar otros tipos de escalamiento de variables o indagar en métodos de clustering que trabajen con otras medidas de disimilaridad, con el fin de incorporar variables cualitativas. Este último aspecto puede ser muy relevante, al permitir la incorporación de la Actividad Económica.

Agradecimientos: Este trabajo fue parcialmente financiado por el Instituto Milenio Sistemas Complejos de Ingeniería.

Referencias

[1] P. BERKHIN, "Survey of clustering data mining techniques", Technical Report, Accrue Software, San Jose CA, 2002.

[2] U. FAYYAD, GREGORY PIATETSKY-SHAPIRO, PADHRAIC SMYTH, "From Data Mining to Knowledge Discovery in Databases", Article, American Association for Artificial Intelligence, 1996, Vol.17, Nº 3, pp. 37-54.

[3] G. FUNG, "A Comprehensive Overview of Basic Clustering Algorithms", June 2001. http://www.cs.wisc.edu/ gfung/clustering.pdf

[4] M. HALKIDI, Y. BATISTAKIS, M. VAZIRGIANNIS, "On clustering validation techniques", Journal Article, Journal of Intelligent Information Systems, Dec. 2001, Vol. 17, pp. 107-145.

[5] J. HANDL, J. KNOWLES, D. B. KELL, "Computational Cluster Validation in Post-genomic Data Anlysis" School of Chemistry, University of Manchester UK, Bioinformatics Review, Vol. 21, Mayo 2005, pp. 3201-3212.

[6] J. HARTIGAN AND M. WONG, "Algorithm AS136: A K-means clustering algorithm", Applied Statistics, Vol. 28, pp. 100-108, 1979.

[7] T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN, "The Elements of Statistical Learning: Datamining, Inference and Prediction", Springer, New York, 2001, Cap. 14, pp. 437-508.

[8] E. PARADIS, "R para Principiantes", Institut des Sciences de lÉvolution, Universit Montpellier II, France, http://cran.r-project.org/doc/contrib/rdebuts-es.pdf.

[9] J. W. SAMMON, JR, "A Nonlinear Mapping for Data Structure Analysis", Transactions on Computers, Mayo 1969, Vol. C-18, Issue 5, pp. 401-409.

[10] SERVICIO DE IMPUESTOS INTERNOS, Formulario Inscripción al Rol único Tributario y/o Declaración de Inicio de Actividades, http://www.sii.cl/formularios/imagen/4415.PDF

[11] SERVICIO DE IMPUESTOS INTERNOS, Formulario Declaración Mensual y Pago Simultáneo de Impuestos (F29), http://www.sii.cl/formularios/anverso-f29.pdf

[12] SERVICIO DE IMPUESTOS INTERNOS, Suplemento Formulario 29, www.sii.cl

[13] B. SILVERMAN, "Density Estimation for Statistics and Data Analysis", Monographs on Statistics and Applied Probability, 1986, Chapman and Hall, London

[14] L.I. SMITH, "A Tutorial on Principal Components Analysis", Febrero 2002, http://www.cs.otago.ac.nz/cosc453/student tutorials/principal-components.pdf

[15] R. TIBSHIRANI, G. WALTHER, T. HASTIE, "Estimating the Number of Clusters in a Dataset via the Gap Statistic", Journal of the Royal Statistical Society: Series B (Statist. Methodol.), Vol. 63, Marzo 2000, pp. 411-423.

[16] L. TORGO, "Data Mining with R: learning by case studies", LIACC-FEP, University of Porto, 22 Mayo 2003, http://www.liac.up.pt/ ltorgo

[17] J.D. VELÁSQUEZ, V. PALADE., "Adaptive web site: A knowledge extraction from web data approach", IOS Press, chapter 3: "Knowledge discovery from web data".

[18] J. VESANTO, "Using SOM in Data Mining", Licentiate´s Thesis, Finland, Abril 2000.

[19] A. WEINGESEL, E. DIMITRIADOU, S. DOLNICAR, "An Examination of Indexes For Determining The Number Of Clusters In Binary Data Sets", Psychometrika, 2002, Vol. 67, Nº 1, pp. 137-160.

[20] I. WITTEN, E. FRANK, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Junio 2005

[21] <http://csnet.otago.ac.nz/cosc453/student-tutorials/principal-components.pdf> [Consulta: Noviembre 2006]

[22] <http://datamining.anu.edu.au/student/math3346-2006/3.4up.pdf> [Consulta: Noviembre 2006).

[23] R PROJECT, <http://www.r-project.org> [Consulta: Junio 2006 a Marzo 2007]

[24] SERVICIO DE IMPUESTOS INTERNOS, <http://www.sii.cl> [Consulta: Junio 2006 a Marzo 2007]

[25] http://www.ir.iit.edu/ñazli/cs422/CS422-Slides/DM-Preprocessing.pdf [Consulta: Noviembre 2006]