Cuadragesimo Sextas Jornadas Argentinas de Informática e Investigación Operativa - 46JAIIO. Sociedad Argentina de Informática e Investigación Operativa, Córdoba, 2017.

Detección de datos anómalos y ruido en bases de datos abiertas para el contralor público, utilizando técnicas de minería de datos.

López-Pablos, Rodrigo y Kuna, Horacio Daniel.

López-Pablos, R., Kuna, H.D. 2017
Detección de datos anómalos y ruido en bases de datos abiertas
para el contralor público, utilizando técnicas de minería de datos
Jornadas Argentinas de Informática - 46JAIIO
Simposio Argentino sobre Tecnología y Sociedad - STS
Universidad Tecnológica Nacional - UTN - FRC

Detección de Datos Anómalos y Ruido en Bases de Datos Abiertas para el Contralor Público, utilizando técnicas de Minería de Datos

Rodrigo López-Pablos[1,2] y Horacio D. Kuna[3]

[1]Escuela de Posgrado, Facultad Regional Buenos Aires,
Universidad Tecnológica Nacional, Argentina
[2]Facultad de Tecnología Computacional y Administración,
Universidad de Informática, Tecnología, Mecánica y Óptica, Rusia
[3]Departamento de Informática, Facultad de Ciencias Exactas, Químicas y Naturales,
Universidad Nacional de Misiones, Posadas, Argentina
{rodrigo.lopezpablos,hkuna}@gmail.com

Resumen. Dentro de la abundante literatura que despliega la minería de datos aplicada al descubrimiento de información en bases de datos, las técnicas de detección de campos anómalos y con ruido han permanecido escasamente empleadas con propósitos dirigidos hacia el contralor cívico y la lucha contra la corrupción; sin embargo, estas pueden ser de suma utilidad en la evaluación de la calidad de bases de datos así como en el descubrimiento de indicios de comportamiento corrupto. En este trabajo, se desarrollan y articulan procedimientos híbridos de detección de datos anómalos y ruido para investigar, experimentar y validar su aplicación en sistemas de declaraciones juradas públicas disponibles actualmente en datos públicos abiertos.

Palabras clave: Datos públicos, Declaraciones Juradas, Bases de datos, Datos anómalos y ruido.

1. Introducción

Los datos públicos (DDPP), como cualquier tipo de datos, pueden encontrarse sujetos a anomalías y ruido como es el caso de cualquier otra base de datos (BBDD); no obstante si estas contienen información acerca de declaraciones juradas de funcionarios públicos de primera línea (DDJJ), en este caso, el descubrimiento de anomalías y ruido podrían implicar tácitamente la existencia de comportamiento corrupto subyacente en datos de funcionarios públicos (FFPP). La existencia de tal posibilidad, como resultado de experimentación con minería de datos, tendrían efectos profundos y de relevancia que podrían terminar impactando en el bienestar

societario dado que la corrupción -tanto pública como privada- condiciona el capital social societario y en última instancia la calidad de vida de las poblaciones.

Los procesos de minería de datos y explotación de la información han sido escasamente usados para la resolución de problemas cívicos vinculados al sector público, procesos de detección de datos anómalos y ruido no son una excepción en cuanto a su potencial; indiscutiblemente, la utilidad de las técnicas de explotación de la información para el descubrimiento de conocimiento en organizaciones civiles que enfrenten el problema de la corrupción, reivindica la utilidad de la minería de datos como herramienta en ciencias sociales, en el monitoreo y la investigación científica [1].

Sucintamente, este trabajo busca demostrar la posibilidad de aplicación efectiva para el contralor civil de algoritmos y técnicas de explotación de la información sobre DDPP en Argentina a pesar de las dificultades a su acceso y la baja calidad de procesamiento de los sistemas disponibles para su acopio. Simultáneamente, se propone la experimentación de la calidad de la bases de DDPP en Argentina, específicamente a lo que hace a DDJJ.

En esta propuesta de investigación se plantean las siguientes hipótesis:

> [i] La experimentación, aplicación y uso de técnicas y procesos de detección de datos anómalo en BBDD de DDPP de DDJJ oficiales constituyen herramientas para encontrar indicios de comportamiento corrupto en la administración pública y lucha contra la corrupción.
> [ii] La evaluación de la calidad de tales BBDD mediante análisis de anomalías y ruido es factible y útil empíricamente.
> [iii] Existen procedimientos de detección de datos anómalos y ruido posibles de ser aplicados considerando el carácter de sistemas de DDJJ disponibles actualmente en acceso abierto.

Posteriormente a esta Sección introductoria, la Sección 2 hace al estado de la cuestión de las técnicas de minería de datos dirigidas al comportamiento corrupto; mientras que en la Sección 3 se estudia la disponibilidad y características de las fuentes de DDPP abiertos y de sistemas de DDJJ patrimoniales existentes en la actualidad. La Sección 4 se presentan los procedimientos de detección de anomalías propuesto, la experimentación en la Sección 5 para finalizar en la Sección 6.

2. Estado de la cuestión

Las técnicas y proceso de minería de datos y explotación de la información han sido escasamente usadas como herramienta cívica de apoyo en la lucha contra la

corrupción desde la esfera ciudadano en el ámbito público; no obstante, encontramos literatura relevante sobre técnicas de minería dirigida a la detección de fraude o corrupción privada, financiera y contable.

Una clasificación completa del uso de técnicas de minería contra la corrupción privada, la cual se manifiesta en la forma de diversos fraudes financieros y contables puede encontrarse, en varios autores [2,3], donde se aprecia una clasificación exhaustiva, comprendiendo fraudes bancarios, tarjetas de crédito, blanqueo o lavado de dinero y fraude hipotecario, casos de corrupción privada compleja fraudes financieros complejos que involucran la falsificación de información corporativa, la especulación financiera, etc. A continuación una clasificación de las técnicas usadas en el campo.

Fig. 1. Clasificación de técnicas de explotación de la información dirigidas al descurbimiento de fraude financiero [2].

La Figura 1 presenta una clasificación de fraudes y técnicas de minería para usadas para combatir el flagelo del fraude financiero [2]; mientras que en la Figura 2 lo hace respecto el fraude contable.

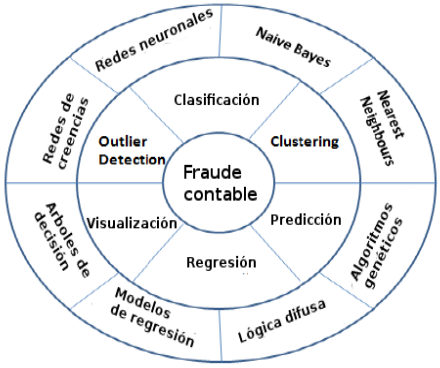

Fig. 2. Clasificación de algoritmos y técnicas de minería de datos utilizados para el descubrimiento de fraude contable o fiscal [3].

De la Figura 2 se pormenoriza una taxonomía abarcativa en la aplicación de minería de datos en la detección de fraude contable [3]. Tácitamente, tales investigaciones sobre minería de datos para el estudio del fraude y corrupción ponen de relieve la subexplotación del potencial de la herramienta en su aplicación cívica y el contralor en administración pública.

De la Figura 1 se advierte el hecho que en el ámbito público, los esfuerzos de utilización de herramientas de minería contra la corrupción han sido subexplotadas, puesto que no encontramos antecedentes ni literatura relevante que así lo demuestre; así mismo, sin desmedro de las restantes técnicas de minería de datos y procesos de explotación de la información, las técnicas de detección de outliers o valores anómalos no ha sido implementada hasta ahora para la solución escenarios de fraude, y por supuesto mucho menos aún, dentro del estudio de la corrupción en el ámbito público.

Fig. 3. Técnicas de explotación de la información dirigidas al descubrimiento de corrupción pública. (Elaboración propia)

Por lo tanto, en la búsqueda y experimentación tanto novedosa como oportuna de procesos desarrollados recientemente para la la detección de campos anómalos y ruido, la detección de anomalías en BBDD públicas de DDJJ, podrían conformar una herramienta útil al descubrimiento de procesos corruptivos que afectan el bienestar y el tejido social.

Los sistemas de DDJJ públicas abiertas son imprescindibles para todo gobierno abierto participativo; los cuales, para ser considerado como tal debe cumplir con los siguientes aspectos.

Tabla 1: Aspectos de los gobiernos abiertos [4].

Condiciones para un gobierno abierto y participativo			
Transparencia fiscal	Acceso a la información	Divulgación patrimonial de FFPP	Compromiso ciudadano

Los sistemas que hacen a todo gobierno abierto son imprescindibles para la prevención de la corrupción, y si bien la transparencia fiscal y la participación ciudadana dependen en mucho de los organismos de contralor y la efectividad de los sistemas educativos en el largo plazo, el acceso a información pública las DDJJ de bienes patrimoniales de FFPP constituyen elementos fundamentales para retroalimentar y fortalecer los círculos virtuosos del autocontralor cívico. Esta retroalimentación puede fortalecerse aún más si se incorporan herramientas de minería de datos para su análisis y comprensión que así lo permitan.

3. Sistemas de declaraciones juradas patrimoniales disponibles como datos públicos abiertos

En Latinoamérica, si bien la intención de constituir gobiernos abiertos ha sido proba, en los hechos, es más lo que se ha realizado desde la ciudadanía, organizaciones civiles y periodísticas que desde los propios Estados, pues aquellos lo han hecho con mayor rapidez y efectividad de manera más acorde a los tiempos digitales actuales. Lo que ha ocurrido en Argentina respecto la accesibilidad a sistemas de DDJJ patrimoniales, representa un caso en este sentido.

En Argentina, los sistemas de DDJJ son sistemas de información que se encuentran regulados por los sistemas o regímenes de DDJJ; los cuales, poseen tres funciones básicas para lo cual fueron implementados [5]:

[i] Control de la evolución patrimonial de los funcionarios de la función pública para prevenir enriquecimiento ilícito y otros delitos de corrupción.
[ii] Detección y prevención de conflictos de intereses e incompatibilidades de la función pública.
[iii] Mecanismo de transparencia y prevención de la corrupción pública.

En sentido preventivo todo régimen de DDJJ de funcionarios públicos representan una herramienta que posibilita: el control del adecuado cumplimiento de las funciones públicas que desempeñan los funcionarios, la prevención del desvío de sus deberes éticos y la corrección de incumplimientos detectados.

Los mismos constan individualmente de documentos donde se expone la variación patrimonial de FFPP durante el desempeño de sus funciones. Sus activos y pasivos, monetarios, muebles e inmuebles, sus antecedentes laborales –en especial aquellas relaciones contractuales mantenidas en forma simultanea en el desempeño del cargo público–, etc. Un sistema de DDJJ de funcionarios también es un sistema de información comprende los siguientes elementos:

Fig. 4. Sistema de DDJJ patrimoniales [4].

Dentro del sistema de la Figura 4, la información recabada en las DDJJ patrimoniales resulta de suma importancia al ser aquellas receptoras y almacenadoras de datos patrimoniales de FFPP con mayor responsabilidad civil y republicana.

3.1 Los sistemas de aplicativos abiertos de DDJJ disponibles en Argentina

Del relevamiento exploratorio de DDPP en Argentina, la experimentación se apoyo en el sistema interactivo DDJJ Abiertas [6] del Diario La Nación en conjunto con las oenegés Directorio Legislativo, Poder Ciudadano y la Asociación Civil por la Libertad y la Justicia, iniciativa vigente desde 2013. Se tomó un total de 1550 DDJJ totales del sitio interactivo, 539 DDJJ son correspondientes a 99 funcionarios del poder ejecutivo, 843 DDJJ correspondientes a 313 funcionarios del poder legislativo, y 168 DDJJ correspondientes a 87 funcionarios del poder judicial; a partir de la versión actualización al 14 de abril de 2015. La estrategia de preparación de los datos para la experimentación se apoyo solamente en los bienes inmuebles de FFPP.

La BD de DDJJ abiertas presentan la siguiente estructura de atributos para cada declaración jurada de funcionario público.

```
dj.funcionario = (ddjj_id, ano, tipo_ddjj, poder, persona_id,
                  nombre, ingreso, cargo, jurisdiccion,
                  cant_acciones, descripcion_del_bien,
                  destino, localidad, nombre_bien_s, origen,
                  pais, porcentaje, provincia, tipo_bien_s,
                  titular_dominio, vinculo, superficiem2,
                  val_decl, valor_patrim)
```

La BD preparada para la experimentación presenta un total de 6627 tuplas con 24 atributos donde, –de 39 atributos iniciales de la BD originaria–, se descartaron 13 al

encontrarse campos parcial o completamente vacíos, inconsistencias nominales en la imputación de los datos, así como atributos que pasaron a ser redundantes *a posteriori* de la homogeneización y estandarización ligados a los tres (3) atributos generados, tal que: `dj.patrimoniales (Gen)=(superficiem2, valor_patrim, val_decl)`. El atributo generado `'val_decl'` puede asumir las siguientes categorías respecto la cualidad del valor declarado de cada inmueble, como se desprende de la Tabla 2.

Tabla 2: Categorías del nuevo atributo generado

Atributo generado Valor declarado (`'val_decl'`)
Valor de Mercado
Valor Fiscal
Valor Subfiscal
No Declara
Sin Datos

Como se expone en la Tabla 2, se categorizó el valor del inmueble declarado por el funcionario `(val_decl)` de acuerdo a su valuación fiscal relativa, pudiendo ser fiscal, subfiscal, de mercado, o no declarándose valor alguno. En los restantes atributos generados, en la búsqueda de homogeneización, previa a la experimención, los valores monetarios patrimoniales en moneda extranjera fueron convertidos a pesos argentinos y actualizados por inflación `(valor_patrim)`, y la superficie inmobiliaria es homogeneizada a metros cuadrados `(superficiem2)`.

4. Metodología híbrida de detección de anomalías y ruido

En estudios recientes, se ha demostrado que los métodos de detección de datos anómalos y con ruido híbridos –aquellos que combinan diferentes algoritmos de distintos enfoques– han revelados como procesos así como la combinación de distintos procedimientos permite detectar outliers con un nivel de confianza mayor al 60% [8].

¿Pero que son los campos anómalos? Estos se definen como un dato que por ser muy diferente a los demás pertenecientes a un mismo conjunto de datos [7]; de esta manera, una base de datos que contiene tales campos, puede considerarse que fue creada por un mecanismo diferente, por lo tanto, en el descubrimiento de tales mecanismos ocultos radica el conocimiento latente en cada base analizada.

A su vez, los métodos híbridos de detección de anomalías, poseen la ventaja de que combinan distintas técnicas y algoritmos para un mismo propósito, por ejemplo: LOF (Local Outlier Factor) y metadatos o LOF y K-Means para BBDD numéricas así como algoritmos de inducción como C4.5, PRISM, Teoría de la Información, RB (Redes Bayesianas); y los algoritmos de clustering LOF, DBSCAN y K-Means para BBDD alfanuméricas con tipos de procedimiento tanto supervisado como no supervisado.

Una descripción sobre la utilización de métodos híbridos –ver la Tabla 3– con enfoques supervisado, no-supervisado y semisupervisado para la detección de outliers y ruido, ya se habían constituido como la mejor alternativa al lograrse la mayor ganancia de información, reducción del espacio de búsqueda y optimización de los procesos [8,9,10]. Investigaciones recientes han determinado que es posible afirmar que la combinación de algoritmos de distinta naturaleza, y también la combinación de procedimientos, permite optimizar el descubrimiento de anomalías [8]; siguiendo ese paradigma, se proponen los dos siguientes subprocedimientos alfanuméricos para el contralor cívico mediante su aplicación en DDJJ.

Tabla 3. Procedimientos según entorno, algoritmos y enfoque [8].

Subprocedimiento	Entorno	Algoritmos y técnicas	Enfoques
III	BBDD alfanuméricas con un atributo objetivo	C4.5, Teoría de la información; LOF	No supervisado, supervisado; Tipo 1 y 2
IV	BBDD alfanuméricas que no contienen un atributo objetivo	LOF; DBSCAN; C4.5; RB; PRISM; K-Means	No supervisado, supervisado; Tipo 1 y 2

El subprocedimiento III, detecta campos de outliers en bases de datos alfanuméricas conteniendo un atributo clase [12,13,8], igualmente al procedimiento IV, el cual también detecta campos en bases alfanumérica solo que sin un atributo target [14,8].

Teniendo conocimiento que las DDJJ se conforman habitualmente por datos alfanuméricos, se descubre la potencialidad de aplicar en su uso los procedimientos híbridos III y IV correspondientes a los procedimientos híbridos de detección de datos anómalos desarrollado recientemente por [8], los cuales se identifican como idóneos por las siguientes razones:

[i] Procedimientos desarrollados recientemente y representan el estado del arte en lo que hace a la detección de campos anómalos y ruido.
[ii] Procedimientos óptimos para la detección de ruido en BBDD alfanuméricas como generalmente se encuentran caracterizados los DDPP.
[iii] Procedimientos de fácil aplicabilidad y ejecución con los programas de minería de datos disponibles actualmente.

Esquemáticamente se describe a continuación las etapas del procedimiento III, primer subprocedimiento propuesto.

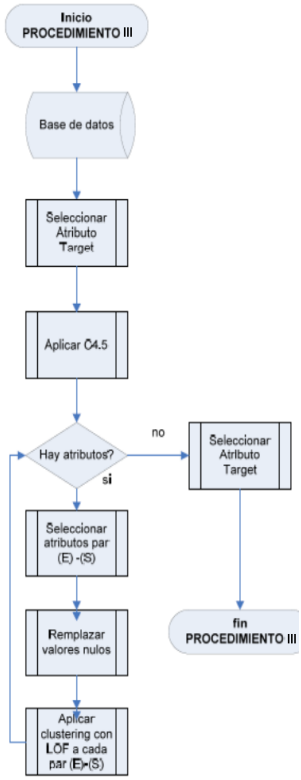

Fig. 5. Procedimiento híbrido alfanumérico III [8].

El procedimiento descripto anteriormente en el diagrama relacional de la Figura 5 anterior, este también puede describirse en seudo-código como sigue inmediatamente.

- Entrada BD
  - Determinar atributo objetivo
  - Aplicar algoritmo **C4.5**
  - Y para cada atributo significativo:
    - Con (E): atributo seleccionado (entrada) y (S) atributo objetivo se arma un bin de entrada-salida (E)+(S).
    - Los datos de entrada (E) son analizados reemplazando los valores nulos por etiquetas.
    - Se aplica **LOF** al bin-conjunto (E)+(S) generando un atributo outlier.

- Se filtran los pares de datos del bin (E)-(S) con valores de outliers ≠ 0, el cual indica la presencia de dato/s anómalo/s al no aportar información y producir ruido en relación al atributo clase.

❍ Salida BD con campos anómalos detectados.

El subprocedimiento híbrido III posee una combinación de distintos tipos enfoques de minería, de aprendizaje supervisado primero –algoritmo C4.5– y aprendizaje no supervisado después al aplicarse LOF. Según experimentación empírica con BBDD reales y artificiales, este procedimiento demostró una efectividad promedio que suele rondar el 90% [8].

Igualmente, el segundo subprocedimiento híbrido de detección de datos anómalos alfanuméricos propuesto, el procedimiento IV, es esquematizado de la siguiente manera en la Figura 6.

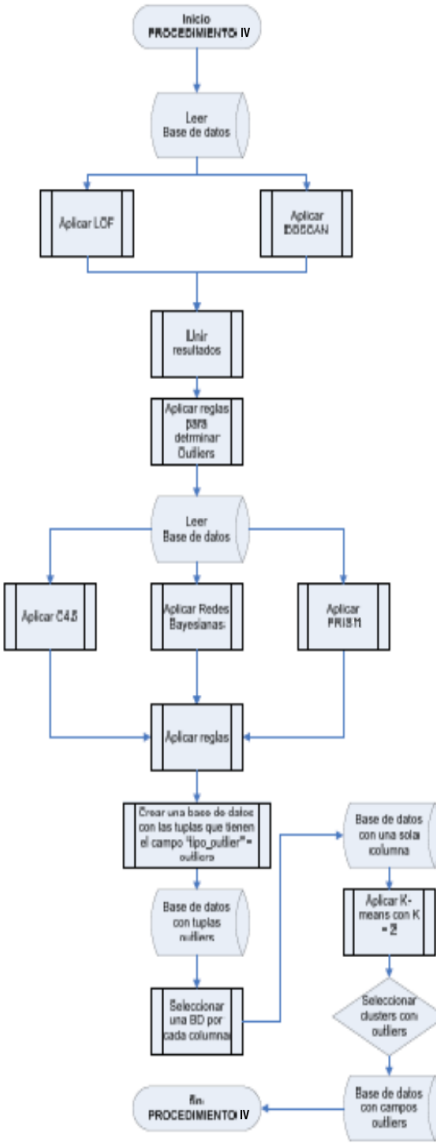

Fig. 6. Procedimiento híbrido alfanumérico IV [8].

Como se observa en la Figura 6, el segundo subprocedimiento propuesto tiene las siguientes etapas, las que pueden ser expresadas en seudo-código como sigue.

- **Entrada** BD(a)
  - Leer BD(a)
    - Aplicar algoritmo **LOF**; agregar atributo `'valor_LOF'` a cada tupla y obtener resultados según **reglas** de determinación de outliers **LOF-DBSCAN**.
    - Aplicar algoritmo **DBSCAN**; agregar atributo `'valor_DBSCAN'` a cada tupla y obtener resultados según **reglas** de determinación de outliers **LOF-DBSCAN**.
  - Unir resultados; agregar atributos `'tipo_outlier'` a cada tupla; obtener resultados según unión de resultados de algoritmos **LOF-DBSCAN**.
  - Leer BD(a) con atributo `'tipo_outlier'`
    - Aplicar algoritmo **C4.5**; determinar valor de atributo clase `'tipo_outlier'` para **C4.5**; obtener resultados.
    - Aplicar algoritmo **RB**; determinar valor de atributo clase `'tipo_outlier'` para **RB**; obtener resultados.
    - Aplicar algoritmo **PRISM**; determinar valor de atributo clase `'tipo_outlier'` para **PRISM**; obtener resultados.
  - Unir resultados
    - Aplicar reglas para la detección de outliers de algoritmos de clasificación.
    - Obtener resultados en cada tupla de atributo clase `'tipo_outlier'` con valor `'limpio'` o `'outlier'`.
    - Crear BD(b), t.q. por cada valor cuyo valor sea `'tipo_outlier'` = `'outlier'`.
    - Clusterizar la primera columna de BD(b), t.q. `'tipo_outlier'` = `'outlier'` con **K-Means** asumiendo **K** = 2
    - Calcular diferencia entre los centroides de los clusters conformados; el cluster más alejado del centroide contiene los campos considerados anómalos.
    - Repetir el procedimiento para cada columna de BD(b) que contenga un valor de `'tipo_outlier'` = `'outlier'`.
- **Salida**, BD(a) con campos anómalos detectados.

Como en el subprocedimiento híbrido III, en el subprocedimiento híbrido IV se combinan distintos enfoque de minería, en una primera fase con aprendizaje no supervisado mediante LOF, DBSCAN primero, y aprendizaje supervisado después, con C4.5, Redes bayesianas y PRISM. En una segunda fase, con una BD con campos outliers producto de la primera fase, se vuelve a un enfoque no supervisado con la aplicación de LOF y K-Means.

BD-DDJJ(a) → BD-DDJJ(b)

(Fase I) (Fase II)

La idoneidad de estos procedimientos híbridos son propuestos como solución tentativa para un caso de contralor civil para la mejora de los DDPP y cívicos de una población. A continuación se despliega una posible resolución y aplicación hipotética de contralor civil en base a los los procedimientos alfanuméricos descriptos, sobre BBDD de DDJJ abiertas previamente tratadas.

Sucintamente ambos subprocedimientos se combinan en un meta procedimiento último para la detección de anomalías como se despliega en la Figura 7 a continuación.

Fig. 7. Propuesta de aplicación de procedimientos híbridos de detección de campos anómalos sobre bases de datos de declaraciones juradas alfanuméricas.[Elaboración propia].

El procedimiento propuesto de la Figura 7, supone la aplicación de ambos subprocedimientos híbridos alfanuméricos de detección de anomalías descriptos *ut supra* sobre las BBDD preparadas de DDJJ para la detección de outliers, la posibilidad de detección de falsos positivos propone la retroalimentación tanto de uno como otro subprocedimiento alfanumérico propuesto. De esta forma, se busca evaluar la calidad de los DDPP implicados en el análisis a primera luz, y en un análisis más profundo *a posteriori* en base a los campos y atributos detectados, los indicios de comportamiento corrupto implícito, existente como output de la información pública procesada.

El meta procedimiento propuesto pueden desprenderse las siguientes etapas en la aplicación de ambos subprocedimientos híbridos de detección de campos anómalos alfanuméricos:

- **Entrada** BD(ddjj)
  - Leer BD(ddjj)

- Aplicar **procedimiento alfanumérico III**.
  - En caso que se encuentren outliers se evalúa la calidad de la BBDD y se exploran indicios de corrupción.
  - En caso contrario se lee la BD(ddjj) nuevamente y se aplica el **procedimiento alfanumérico IV**.
  - Leer BD(ddjj)
- Aplicar **procedimiento alfanumérico IV**.
  - En caso que se encuentren outliers se evalúa la calidad de la BBDD y se exploran indicios de corrupción.
  - En caso contrario se lee la BD(ddjj) nuevamente y se aplica el **procedimiento alfanumérico III**.

❍ **Salida**, calidad de la BD(ddjj) evaluada, con o sin indicios de corrupción detectados.

Con el procedimiento descripto se pasa al abordaje de la experimentación.

5. Resultados de la experimentación

La experimentación contempla la experimentación en ambos subprocesos híbridos de detección de campos anómalos en BBDD de DDJJ alfanuméricas.

5.1. Validación del subprocedimiento alfanumérico III

De la ejecución experimental del primer procedimiento alfanumérico, en su fase no automática a `'val_decl'` como atributo clase del algoritmo de inducción C4.5 se obtienen los siguientes atributos como significativos al lograr la mayor ganancia informativa (Tabla 4).

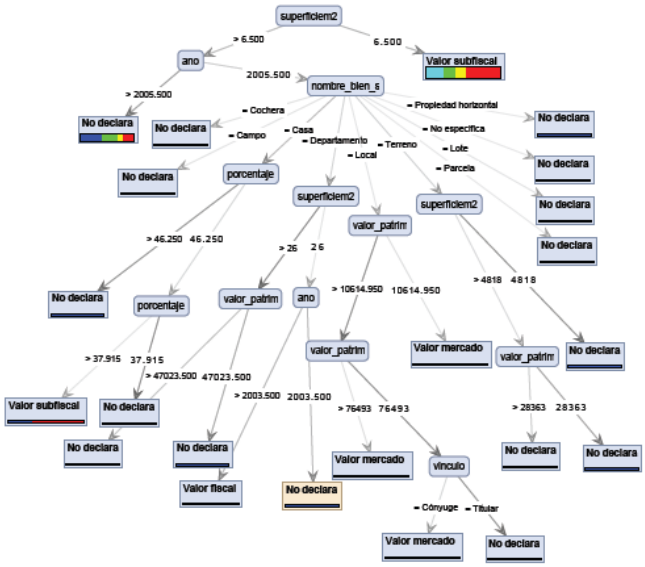

Fig. 8. Árbol de inducción resultado de la aplicación del algoritmo C4.5.

Del análisis de inducción desplegado en la Figura 8, se desprenden los siguientes atributos significativos para explicar la declaración patrimonial de bienes inmuebles de acuerdo a su valor relativo presentes en la Tabla 4 desplegada como sigue.

Tabla 4.
Atributos significativos detectados.

Atributos significativos
superficiem2
ano
nombre_bien_s
porcentaje
val_patrim
vinculo

Del primer análisis de inducción con atributo clase –subprocedimiento III–, mediante algoritmo C4.5, también se observaron las siguientes reglas:

[i] Los inmuebles menores a 6500m^2 tienden a ser declarados a valor subfiscal.
[ii] Los inmuebles mayores a 6500m^2 –entre mediados de 2005 y 2012– tienden a no declararse, ocurriendo lo mismo con cocheras, campos, terrenos, parcelas, lotes y propiedades horizontales sin especificación mayores a 6500m2 y anteriores a 2005.
[iii] Cuando el funcionario posee una casa, y una participación accionaria superior al 46,3% [100%; 46,3%) o inferior al 37,9 (37,9% ; 0%] prefiere no declarar su valuación, pero tiende a declararla subfiscalmente cuando posee un rango accionario entre el 46% y el 38% (46,3%; 37,9%).
[iv] Si el funcionario posee un departamento, entre mediados de 2003 y mediados de 2005, con superficie menor o igual a 26m^2, tiende a ser declarado a su valuación fiscal, mientras no se declara cuando es anterior a 2003. No obstante, los departamentos mayores a 26m^2 simplemente tienden a no declararse.
[v] Los locales comerciales de los FFPP que poseen una valuación mayor a $76493 [∞, $76493) o inferiores a $10615 ($10615, 0] tienden a ser declarados a su valor de mercado; a excepción de los locales con rango patrimonial entre [$76493, $10615], dado que en este caso también se tenderá a ser declarado a su valor de mercado, si y solo si, el inmueble se encuentra inscripto a nombre de su cónyuge, y no precisamente a nombre del oficial político, caso en el cual se tenderá a no declararlo.

El flujo de minería procede tomando los atributos significativos de la Tabla 4 hallados con C4.5 sobre los cuales se elaboraron 6 bins de entrada-salida simulando un sistema de información (Tabla 5); para *a posteriori*, ejecutarlos en la fase automática no supervisada del procedimiento III con algoritmo LOF, obteniendo los siguientes resultados de la Tabla 5.

Tabla 5. Bins de Entrada-Salida con outliers detectados.

Bins Entrada-(Salida)	**Outliers detectados**	**Bins anómalos sospechosos Media o Moda(Moda)**
superficiem2-(val_decl)	968 (∞)	38733.15(No declara)
ano-(val_decl)	122 (∞)	2001(Mercado)
nombre_bien_s-(val_decl)	209 (∞)	Prop. Horizontal(Fiscal)
porcentaje-(val_decl)	252 (∞)	29.18528(Sin datos)
val_patrim-(val_decl)	1130 (∞)	146528.3(Subfiscal)
vinculo-(val_decl)	30 (∞)	Conviviente(Subfiscal)

Donde los atributos superficie patrimonial, nombre del bien y porcentaje accionario presentaron la mayor cantidad de anomalías así como el año de la DDJJ, el valor patrimonial del inmueble, y el vinculo familiar del funcionario declarante siendo todos ellos significativos para explicar el tipo de valuación patrimonial de los inmuebles declarados. A partir de la teoría de la información [15] se tiene que a mayor cantidad de anomalías le corresponde un mayor ruido y entropía en el sistema de DDJJ estudiado.

5.2. Validación del subprocedimiento alfanumérico IV

La ejecución del subprocedimiento IV, sin atributo target, en la primera fase automática se aplican LOF-DBSCAN siguiendo reglas de determinación de outlier [8] y unión de algoritmos de clasificación C4.5-RB-PRISM no automática, donde solo como resultado parcial y preliminar a la ejecución completa del subprocedimiento IV, se detectaron 2531 tuplas anómalas, *i.e.*, un 38,19% del total de tuplas, como se desprende de la Tabla 6.

Tabla 6. Outliers detectados por unión de algoritmos de clasificación.

Algoritmo	Outliers detectados
UNIÓN RB-C4.5-PRISM	2531

A partir de donde se conforma una BD de outliers para la segunda fase en la cual se transforma *a posteriori* para la aplicación del algoritmo K-Means en dos agrupamientos (n=2), observándose los siguientes resultados para cada atributo como se presenta en la Tabla VII.

Tabla 7. Distancia del centroide para cada atributo.

Atributo(id)	Valor distancia Cluster 0	Valor distancia Cluster 1
ddjj_id(1)	**1.841**	
ano(2)	**1.518**	
tipo_ddjj(3)		1.302
poder(4)		1.232
persona_id(5)		1.136
nombre(6)		1.063
ingreso(7)		1.062
cargo(8)		1.035
jurisdiccion(9)		1.009
cant_acciones(10)		0.937
descripcion_del_bien(11)		0.926
destino(12)		0.942
localidad(13)		1.042
nombre_bien_s(14)		0.998
origen(15)		1.054
pais(16)		1.003
porcentaje(17)		1.071
provincia(18)		1.104
tipo_bien_s(19)		1.050
titular_dominio(20)		1.076
vinculo(21)		1.038
superficiem2(22)	**2.033**	
val_decl(23)		1.095
valor_patrim(24)		1.237
outlier(25) (Transpuesto BD(b))		1.326

De la Tabla 7 superior se presentan los valores promedios de las distancias desde el centroide para cada atributo en cada agrupamiento correspondiente al K-Means, nótese que aquellos atributos sospechosos de contener anomalías poseen una mayor distancia respecto su centroide y son agrupados en `'cluster_0'` siendo estos, el identificador de la declaración jurada (`ddjj_id`), el año (`ano`) de presentación y la superficie cuadrada del inmueble (`superficiem2`), al poseer estos últimos las mayores distancias respecto sus centroides.

Si bien al ejecutarse el algoritmo K-Means, el centroide más alejado contuvo los atributo superficie patrimonial, el atributo más alejado así como el más sospechoso de contener campos anómalos, el atributo valor patrimonial (`valor_patrim`) a pesar de conformar parte del cluster de atributos sin anomalías, también presenta una distancia relativa importante.

6. Conclusiones

En este trabajo se constituyeron y combinaron procesos híbridos para la detección de outliers y ruido configurados para trabajar con BBDD alfanuméricas en BBDD de DDJJ de bienes inmuebles, como caso inédito de uso de técnicas y procesos de minería de datos como herramienta contra la corrupción pública, poniéndose de relieve la potencialidad de los procesos de detección de anomalías a la hora de evaluar la calidad de las BBDD públicas por un lado, y el descubrimiento de información sospechosamente portadora de comportamiento corrupto por el otro.

Los atributos superficie patrimonial, porcentaje accionario, nombre del bien y año son los que aportaron más entropía al sistema de DDJJ comprometiendo así la calidad de la BD; dada su baja densidad relativa, la superficie del inmueble es conjuntamente con el de valor patrimonial conforman los atributos más sospechoso de contener datos anómalos. De esta manera las anomalías detectadas podrían revelar indicios de comportamiento corrupto respecto la valuación fiscal inmobiliaria declarada por el funcionario, puesto que las características del inmueble podrían condicionar la valuación patrimonial del mismo a su valor fiscal, subfiscal, o simplemente evitando declarar su valor.

Futuros trabajos de investigación podrían contemplar la elaboración de variaciones algorítmicas en los procedimientos alfanuméricos propuestos; no descartándose la aplicación de variantes a los procesos alfanuméricos aquí presentados.

Desde la crisis informacional contemporánea que hace a la transición de los sistemas de representación republicanos tradicionales hacia los digitales, este trabajo contribuye con una herramienta aplicativa más para el contralor del Estado y la participación ciudadana. Lamentablemente no se cuentan con datos abiertos de bases de datos capaces de evaluar la corrupción privada, la que al pertenecer a una misma

sociedad, puede presumirse análogamente severa a la corrupción existente en la esfera pública.

Desde la economía de la distribución, la detección de outliers en atributos económicos puede interpretarse como la presencia de factores que insinúan la existencia de desigualdad patrimonial extrema entre los funcionarios. En primera instancia, el conocimiento de tales concentraciones de riqueza pueden llegar a amenazar la estabilidad societaria, pues cuando el funcionario público no reconoce la dinámica de su propio acervo patrimonial como una extensión de su labor cívica, da lugar a inconsistencias que tarde o temprano llevan a consecuencias severas sobre el capital social, al sopesarse desde la ciudadanía la ejemplaridad ideal del funcionario respecto la comunidad representada.

Referencias

1. Ransom J., Replicating Data Mining Techniques for Development: A Case of Study of Corruption, Lund University, Master Thesis, Master of Science in International Development and Management, 2013., http://lup.lub.lu.se/record/3798253/file/3910587.pdf
2. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems, 50(3), pp. 559-569, 2011.
3. Sowjanya, S., Jyotsna G.: Application of Data Mining Techniques for Financial Accounting Frad Detection Scheme, International Journal of Advanced Research in Computer Science and Software Engineering, Noviembre, 3(11), pp. 717-724, 2013.
4. Open Government Partnership, OGP: Articles of Government, June, 2012.
5. Gómez N., Bello M. A.: Ética, transparencia y lucha contra la corrupcion en la administracion publica, Manual para el ejercicio de la funcion publica, CABA: 1ra ed.: Oficina Anticorrupción, Ministerio de Justicia y Derechos Humanos de la Nación, Mayo, 2009. http://www.anticorrupcion.gov.ar/documentos/Libro%20SICEP%202da%20parte.pdf
6. DD.JJ. Abiertas, LNData, CABA: Actualizado al 13/1/2014, 2015, http://interactivos.lanacion.com.ar/declaraciones-juradas/
7. Hawkins, D. M., Identification of outliers, London: Chapman and Hall., 11, 1980.
8. Kuna H.: Procedimientos de explotación de la información para la identificación de datos faltantes con ruido e incosistentes, Tesis doctoral, Universidad de Málaga, Marzo, 2014.
9. Kuna, H., García Martínez, R., Villatoro, F.: Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información, Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, 5, pp. 39–44, 2009.
10. Kuna, H., García-Martínez, R. Villatoro, F.: Pattern Discovery in University Students Desertion Based on Data Mining, In Advances and Applications in Statistical Sciences Journal, 2(2), pp. 275–286, 2010.
11. Kuna, H., Pautsch, G., Rey, M., Cuba, C., Rambo, A., Caballero, S., Steinhilber, A., García-Martínez, R., Villatoro, F.: “Avances en procedimientos de la explotación de información

con algoritmos basados en la densidad para la identificación de outliers en bases de datos," Proceedings XIII Workshop de Investigadores en Ciencias de la Computación. Artículo 3745, 2011.

12. Kuna, H. , Pautsch, G., Rey, M., Cuba, C., Rambo, A., Caballero, S., García-Martínez, R., Villatoro, F.: Comparación de la efectividad de procedimientos de la explotación de información para la identificación de outliers en bases de datos. Proceedings del XIV Workshop de Investigadores en Ciencias de la Computación, 296--300 (2012b).
13. Kuna, H., Pautsch, G., Rambo, A., Rey, M., Cortes, J., Rolón, S.: Procedimiento de explotación de información para la identificación de campos anómalos en base de datos alfanuméricas. Revista Latinoamericana de Ingeniería de Software, 1(3): 102--106 (2013b).
14. Kuna, H., García-Martínez, R., Villatoro, F.: Outlier detection in audit logs for application systems. Information Systems (2014).
15. Ferreyra M.: Powerhouse: Data Mining usando Teoría de la información, (2007).