Exploración de la violencia sexual en la ciudad de Bogotá: una aplicación de técnicas de minería de datos

Exploration of sexual violence in the city of Bogota: application of a data mining technique

Exploração da violência sexual na cidade de Bogotá: uma aplicação das técnicas da mineração dos dados

FECHA DE RECEPCIÓN: 2011/03/15 FECHA DE ACEPTACIÓN: 2011/12/15

Nelson J. Garnica
Sociólogo.
Consultor, Asesor, Fondo de Vigilancia y Seguridad de Bogotá, Bogotá, D. C., Colombia.
nelson.garnica@gmil.com

Ángela Marcela Olaya-Murillo
Socióloga.
Asesora, Fondo de Vigilancia y Seguridad de Bogotá, Bogotá, D. C., Colombia.
aolayamurillo@gmail.com

Resumen

Este estudio ofrece una aproximación a la violencia sexual utilizando datos de fuente secundaria y aplicando algunas técnicas de minería de datos. La fuente de datos utilizada es el Instituto Nacional de Medicina Legal y Ciencias Forenses, y los algoritmos aplicados son Selección por Características, C5.0 y K-Means. Antes de la aplicación de dichas técnicas se hace una aproximación teórica a la violencia sexual, para apreciar la forma como se ha abordado este tipo de violencia y la manera como se ha analizado. Seguidamente se realiza la evaluación de la calidad de los datos y se aplican algunos tratamientos para su mejoramiento. Una vez se llega a un dataset adecuado para el procesamiento y análisis, se aplican técnicas de minería de datos y se establece como variable objetivo o respuesta la relación del presunto agresor con la víctima.

Las salidas que ofrece el procesamiento llevan a un análisis que establece como centro los niveles de proximidad con la víctima y cuestiona los análisis basados en la estructura de parentesco tradicional, al mismo tiempo que valida la distinción que establece una clasificación de la violencia sexual entre abuso sexual y asalto sexual. Los análisis del ejercicio de minería de datos

permiten plantear claramente la configuración de dos clusters a los que se les puede señalar con dicha clasificación. Estos están acompañados de un tercero que, si bien no está muy definido, empieza a dibujarse. Los tres clusters se han llamado violencia sexual en situación de incesto, violencia sexual en situación de anonimia y violencia sexual en situación de estructura familiar. Se termina con algunas sugerencias en procura del mejoramiento de la calidad de los datos y se plantean las posibilidades que este tipo de análisis tiene al intentar dar respuesta a la conflictividad, la violencia y el delito.

Palabras clave

Delitos sexuales, datos cuantitativos, víctima, familia, control social (fuente: Tesauro de política criminal latinoamericana - ILANUD).

Abstract

This study offers an approach to sexual violence by using secondary source data and applying some data mining techniques. The data source used is the 'Instituto Nacional de Medicina Legal y Ciencias Forenses' (National Institute of Legal Medicine and Forensic Sciences), and the algorithms applied are Selection by Characteristics, C5.0, and K-Means.

Prior to applying these techniques, a theoretical approximation to sexual violence is made in order to appreciate how this kind of violence has been approached and analyzed. Subsequently, data quality is assessed and some improvement treatments are applied. Upon having reached a proper dataset for processing and analysis, data mining techniques are applied, and the relationship of the alleged aggressor to the victim is established as a variable objective or answer.

The issues or solutions offered by the above data processing lead to an analysis which establishes as a core the levels of proximity with the victim, and questions those studies based on the traditional kinship structure, while it simultaneously validates the distinction that establishes a sexual violence rating between abuse and assault. Analyses of the data mining exercise facilitate a clear statement of the configuration of two clusters that can be pointed at with said classification. They are accompanied by a third one that, although not well defined yet, begins to appear. The three clusters have been designated as sexual violence in an incest situation, sexual violence in an anonymity situation, and sexual violence in a family structure situation. Finally, some suggestions are given in seeking to improve data quality, while the opportunities this type of analysis opens at attempting to give an answer to conflictivity, violence and crime are outlined.

Key words

Sexual offenses, quantitative data, victim, family, social control (Source: Tesauro de política criminal latinoamericana - ILANUD).

Resumo

Este estudo oferece uma aproximação à violência sexual usando dados da fonte secundária e aplicando algumas técnicas da mineração dos dados. A fonte dos dados usada é ao Instituto Nacional de Medicina Legal e Ciências Forenses, e os algoritmos aplicados são Seleção por Característica, C5.0 e K-Means. Antes que a aplicação destas técnicas, uma aproximação teórica à violência sexual é feita, para apreciar o jeito como este tipo de violência é abordado e a maneira como foi analisado. A avaliação da qualidade dos dados é feita em seguida e alguns tratamentos para sua melhoria são aplicados. Uma vez que o dataset adequado é alcançado para o processamento e a análise, as técnicas da mineração dos dados são aplicadas e a relação do suposto agressor com a vítima é estabelecida com a variável alvo ou a resposta.

As saídas que fornece o processamento leva a uma análise que estabelece como o centro os níveis da proximidade com a vítima e questiona as análises baseadas na estrutura do parentesco tradicional, ao mesmo tempo em que valida à distinção que estabelece uma classificação da violência sexual entre o abuso sexual e o assalto sexual. As análises do exercício da mineração dos dados permitem expor claramente a configuração de dois clusters aos quais é possível assinalar com esta classificação. Estes são acompanhados de um terceiro que, embora não esteja bem definido, começam a extrair-se. Os três clusters foram chamados violência sexual na situação do incesto, violência sexual na situação de anonímia e a violência sexual na situação da estrutura familiar. Termina com algumas sugestões nas tentativas da melhoria da qualidade dos dados e expor as possibilidades que este tipo de análise tem ao tentar dar resposta aos conflitos, à violência e ao crime.

Palavras-chave

Crimes sexual, dados quantitativos, vítima, família, controle social (fonte: Tesauro de política criminal latinoamericana - ILANUD).

Introducción

Para la realización de un estudio sobre violencia sexual se requiere precisión conceptual del constructo que se pretende analizar. Es pertinente señalar algunos elementos de carácter conceptual que indiquen lo que se entiende por dicho fenómeno y, también, la oposición de la sociedad a determinado comportamiento expresado en los años de pena para cada tipo de delito y consignado en el Código Penal colombiano. A partir de estos presupuestos conceptuales se aborda el mecanismo de análisis, centrado en los datos que pretenden contribuir a la compresión del fenómeno.

Violencia sexual: algunas aproximaciones conceptuales y normativas

La Organización Mundial de la Salud[1] define la violencia como "el *uso intencional* de la fuerza o el *poder* físico, de hecho o como amenaza, contra uno mismo, otra persona o un grupo o comunidad, que cause o tenga muchas probabilidades de causar lesiones, muerte, daños psicológicos, trastornos del desarrollo o privaciones" (2003, p. 5). Un marco para comprender los tipos de violencia se obtiene captando la naturaleza de los actos, que pueden ser físicos, sexuales, psíquicos, privaciones o descuido; el entorno; la relación entre el agresor y la víctima, y los posibles motivos de la violencia. Esta puede ser autoinfligida, interpersonal o colectiva.

Desde un punto de vista la violencia sexual se ha definido como las relaciones sexuales sin consentimiento a través de la manipulación e imposición física o psíquica, donde el victimario impone una relación sexual o acto con connotación sexual no deseado mediante coacción, intimidación o sometimiento a un estado de indefensión (Pinzón, 2009).

En Colombia el acceso carnal y los actos sexuales violentos, se contemplan dentro del Código Penal, tipificados en la legislación sobre violencia sexual y actos sexuales abusivos, definidos como un delito contra la libertad, integridad y formación sexual. El artículo 205 define el acceso carnal violento como "el que realice acceso carnal con otra persona mediante violencia". El artículo 206 define el acto sexual violento como "el que realice en otra persona acto sexual diverso al acceso carnal mediante violencia". En el artículo 207 se define el acceso carnal o acto sexual en persona puesta en incapacidad de resistir y que puede estar en condiciones de inferioridad psíquica o trastorno mental. En el artículo 209 se definen los actos sexuales con menor de catorce años como "el que realizare actos sexuales diversos del acceso carnal con persona menor de catorce años o en su presencia, o la induzca a prácticas sexuales".

Desde una perspectiva médico-social, el Instituto Nacional de Medicina Legal y Ciencias Forenses[2] define abuso sexual como "el contacto o interacción entre un menor y un adulto, en el que el menor de edad es utilizado para la satisfacción sexual del adulto o de terceros, desconociéndose su nivel de desarrollo psicosexual" (González, J., 2007, p. 144). Se diferencia el abuso sexual del asalto sexual, que hace referencia a la "modalidad específica de agresión caracterizada por actos de violencia física y/o psicológica perpetrado sobre una víctima de cualquier edad o sexo, por el cual persigue un propósito sexual definido" (*op. cit.*).

Desde la comprensión de los diferentes tipos de violencia, la OMS define como violencia sexual "todo acto sexual, la tentativa de consumar un acto sexual, los comentarios o insinuaciones sexuales no

1 De aquí en adelante OMS, por sus siglas en español.

2 De aquí en adelante INML-CF, por sus siglas.

deseados, o las acciones para comercializar o utilizar de cualquier otro modo la sexualidad de una persona mediante coacción[3] por otra persona, independientemente de la relación de esta con la víctima, en cualquier ámbito, incluidos el hogar y el lugar de trabajo". La violación se entiende como la "penetración forzada físicamente o empleando otros medios de coacción, por más leves que sean, de la vulva o el ano, usando un pene, otras partes corporales o un objeto" (*op. cit.*, p. 161).

La violencia sexual es uno de los delitos que impacta de forma muy negativa el bienestar de la sociedad, y de manera dramática la existencia de los individuos afectados directa o indirectamente, como víctimas o familiares de las víctimas. Esta puede producirse en circunstancias y ámbitos distintos. Según el INML-CF, la vivienda se presenta como escenario de mayor riesgo, en relación con el total de los casos. El Centro de Estudio y Análisis en Convivencia y Seguridad Ciudadana[4], sobre violencia intrafamiliar y abuso sexual, de acuerdo con los registros del INML-CF, identifica que si bien el delito sexual también es perpetrado en otro tipo de escenarios como en un vehículo, parque y/o bosque, centro educativo, hotel/motel, u otro lugar público, este se perpetra principalmente en la vivienda, tanto para las mujeres como para los hombres (2008). Es decir "puertas para adentro", lo cual dificulta su prevención, cuidado, atención y protección, y la posibilidad de tratamiento y rehabilitación del hecho que puede afectar el desarrollo psicosexual y la calidad de vida de la persona agredida y su entorno.

En Bogotá se encuentra que la violencia sexual puede ser perpetrada por agresores dudosos o desconocidos, que no son familiares de la víctima y/o grupos de delincuencia común, principalmente. No obstante, la violencia sexual se presenta con mayor frecuencia cuando el agresor es familiar de la víctima, puede ser padre o madre, padrastro o madrastra, tío, primo, abuelo, hermano, ex esposo, esposo o compañero. En el caso en que el presunto agresor no es familiar de la víctima, con mayor frecuencia los perpetradores son amigos, vecinos, conocidos, novios, encargados del menor, profesores, entre otros (*op. cit.*).

Sobre la temporalidad del hecho se pueden identificar varios aspectos. Los hechos se incrementan sobre todo los fines de semana. La mayoría de los casos ocurren entre las 12:00 del mediodía y las 6:00 de la tarde, lo cual se relaciona con la hora de salida cuando se estudia en la jornada de la mañana. En las zonas urbanas se presentan la mayoría de los hechos, donde se observan fenómenos de despersonalización, aislamiento, barreras por el establecimiento de vínculos y de redes sociales, unidos en algunos casos con el hacinamiento poblacional (*op. cit.*, 2007, p. 145).

3 Puede ser el uso de fuerza física, intimidación psíquica, extorsión, amenazas, o cuando la persona agredida no está en capacidad de dar su consentimiento.

4 De aquí en adelante CEACSC, por sus siglas.

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

Un último aspecto por señalar de acuerdo con cifras en relación con la edad de la víctima y el sexo "por cada niño, se atienden 6 niñas y por cada persona adulta se atienden 3,1 menores de edad" (*op. cit.*). Lo anterior se puede relacionar con el nivel de escolaridad, la mayoría de casos corresponden a niños con primaria y secundaria incompleta, siendo los grupos más afectados los estudiantes y quienes se dedican al hogar.

Se hace evidente, entonces, lo pertinente de indagar por los aspectos estructurales que llevan a que se comentan actos de violencia, específicamente de violencia sexual, que por su impacto social y sobre los imaginarios referentes a la seguridad civil, dejan expuesto el problema del riesgo como producto de la vulnerabilidad, la amenaza real y potencial.

El análisis de bases de datos en la prevención de la violencia

Con el fin de determinar las causas de la criminalidad y movilizar los diferentes componentes de la sociedad, cada vez más procesos tratan de apoyar sus acciones de prevención sobre una base científica. Existe un desarrollo a escala internacional interesado en las políticas de prevención fundadas en datos más confiables. La prevención depende de las intervenciones sobre cuatro enfoques señalados por el Centro Internacional para la Prevención de la Criminalidad (2008).

La *prevención social* prioriza el bienestar y la cohesión social a través de acciones en materia de salud, educación, desarrollo económico y social; por la movilización de los miembros de la comunidad se da la *prevención comunitaria*; *la prevención situacional* o de las situaciones propias al delito, y *la prevención de la reincidencia*. La prevención basada en el conocimiento se relaciona con la dirección de acciones que ayuden al ejercicio de interpretación y difusión de los datos, involucrados con el tema de la percepción de seguridad, sobre los cuales los medios masivos de comunicación invariablemente intervienen a favor o en contra (*op. cit.*).

Política pública basada en el conocimiento

Los estudios estadísticos en las sociedades contemporáneas se han convertido en un recurso para explicar las condiciones y dinámicas de cambio, por medio de la apreciación objetiva del comportamiento agregado de los registros, para estimar la transformación, acercamiento o distanciamiento a metas compartidas por los individuos en el seno de la sociedad, con el fin de estimar el estado de los fenómenos tanto deseables como indeseables; para este caso la condición de seguridad y el comportamiento de la violencia sexual que se presentan en la ciudad de Bogotá.

El análisis de los registros de delitos se fundamenta en la prevención de la conflictividad, la violencia y el delito, para lograr unas intervenciones de prevención efectivas. Los mecanismos, tecnologías y técnicas de análisis del delito desempeñan una función en el logro de objetivos, seguimiento eficiente y divulgación de la información, por medio de la extracción de información relevante, para la generación de conocimiento, cambios en la conducta y la percepción de los individuos.

Método

La conjugación de los últimos desarrollos en computación y tecnologías de la información ha llevado a desarrollos sobre procesamiento de datos y extracción de conocimiento, con técnicas y herramientas robustas que permiten la manipulación y explotación de bases de datos. Estos

desarrollos se conocen generalmente como Descubrimiento de Conocimiento en Bases de Datos, proceso que cubre un espectro amplio de tareas y técnicas que van desde la obtención de acceso a las fuentes de información; depuración y puesta a punto de los datos; validación y selección; aplicación de técnicas de minería de datos, hasta generación y preparación de divulgación de los resultados.

La depuración de la información de violencia sexual de la fuente se realizó con el fin de perfilar los datos, diseñando estrategias adecuadas para manejar ruido, valores incompletos, valores fuera de rango, valores inconsistentes y en blanco. Para el mejoramiento de calidad de los datos, se realizó un procesamiento en dos fases: una depuración manual para corregir errores de entrada y la construcción de algunos atributos.

En la segunda fase se aplicaron mecanismos automáticos para la determinación de la calidad a nivel de campos o variables, y de registros o casos. Los algoritmos de minería de datos aplicados fueron Selección por Características y Detección de Anomalías. El informe de calidad de los datos permite su consolidación por medio de la selección de los campos de interés, depuración de registros en busca de completitud y consistencia, y modificación de las variables de los campos en función de los algoritmos a utilizar.

En este marco se propone un abordaje metodológico consistente con la tradición científica en el área estadística y con metodologías utilizadas a nivel mundial en Minería de Datos (Chen et al., 2004; Zeleznikow, 2005). La minería de datos, así como el descubrimiento de conocimiento en los datos[5], integran principios teóricos y desarrollos metodológicos provenientes de la estadística, el aprendizaje automático, la inteligencia artificial, la visualización de datos y la teoría de bases de datos.

La necesidad de disponer de una mayor cantidad de elementos para establecer políticas de inteligencia criminal, obliga a evolucionar en el procesamiento y análisis de la información. Si bien los métodos de investigación son complementarios, la estadística plantea hipótesis a ser validadas con los datos disponibles, y la minería de datos persigue el descubrimiento de patrones de comportamiento social, no previstos desde la estadística. No obstante, las exploraciones aquí realizadas no agotan las posibles herramientas de minería de datos, solo se utilizan unas cuantas de las disponibles y el propósito de fondo es señalar las potencialidades que dichos desarrollos ofrecen en la explotación de bases de datos con fines de extracción de conocimiento.

La cantidad de información y variables intervinientes en los registros de violencia sexual justifican el uso de herramientas complementarias a la estadística convencional, para determinar relaciones multivariantes subyacentes. La minería de datos es un proceso de extracción de información y conocimiento no trivial en grandes volúmenes de datos (Kantardzic, 2002), cuya aplicación a la inteligencia criminal se ha constituido en un campo relativamente nuevo, con gran impulso en los últimos años en Estados Unidos (Chen et al., 2004), para generar información como patrones, asociaciones, cambios, anomalías y estructuras significativas (Ochoa, 2004), que no son fáciles de observar de manera directa.

En el estudio se aplicaron algunas herramientas de minería de datos al análisis sobre violencia sexual en la ciudad de Bogotá, en búsqueda de conocimiento nuevo y valioso y/o validar conocimientos adquiridos hasta el momento. Se espera realizar una contribución para la modernización de las

5 De aquí en adelante KDD, por sus siglas en inglés: *Knowledge Discovery in Database*.

prácticas sobre el tratamiento y análisis de datos e información sobre delitos a nivel local, permitiendo una evaluación de la calidad de estos y evidenciando la capacidad de las técnicas de minería de datos para extraer conocimiento de grandes volúmenes de información. En general, se propone pasar del *dato* sobre el delito y la *información* sobre este, al *conocimiento* del comportamiento social de la violencia sexual en la ciudad de Bogotá, con información recopilada a nivel distrital.

Se realizó un análisis exhaustivo de la información recolectada, determinando relaciones multivariantes subyacentes y extrayendo conclusiones para constituirse en un valor agregado, basado en conocimiento científico sobre el comportamiento de la violencia sexual. El Sistema Unificado de Información de Violencia y Delincuencia, hoy en día el CEACSC, analiza la información del Distrito por medio de un análisis estadístico, expresado en tasas por 100.000 habitantes y número de eventos. Sin embargo, es fundamental encontrar patrones vinculados con el tipo de presunto agresor, variables situacionales y características sociodemográficas de la víctima, que permitan generar nuevo conocimiento sobre la problemática y/o validar el adquirido hasta el momento.

Las técnicas y herramientas de minería de datos sobre la información disponible del INML-CF, mejoraron la calidad de los obtenidos sobre violencia sexual y se logró la identificación de patrones. Lo anterior permitió, por un lado, la aplicación del algoritmo *K-Means* para agrupar los hechos según su similitud en grupos o *clusters* distintos, y por otro lado, el uso del algoritmo de inducción C5.0 permitió identificar reglas de pertenencia a cada uno de los grupos o *clusters*.

Fuentes de información para el análisis

En el ámbito nacional, las fuentes de información criminal son el Instituto Nacional de Medicina Legal y Ciencias Forenses, el Centro de Investigaciones Criminológicas de la Policía Metropolitana de Bogotá y la Fiscalía General de la Nación. Para este estudio se tomó el *dataset* sexológicos del INML-CF, que registra la violencia sexual para el año 2007 en la ciudad de Bogotá. El INML-CF es la organización pública de referencia técnico-científica, que produce información recolectada, procesada, analizada y divulgada a través de la actividad forense y, particularmente, es la que se encuentra a la entrada de todo proceso de investigación científica que busca determinar las circunstancias en las que ocurren los hechos punibles y los delitos.

Se entiende la información sobre delitos como toda aquella resultante de un presunto delito o hecho punible y sus componentes, que sea relevante para la toma de decisiones *a posteriori*, ya sea en la prevención, detección y esclarecimiento del delito, como en el proceso de delincuentes o criminales; la mejora de procesos judiciales; la creación o reforma de leyes, y la intervención a víctimas. Se entiende como caso toda víctima de una lesión de causa externa fatal o no fatal, que es de conocimiento del Instituto (González, J., 2007).

Si bien las cifras solo corresponden a una parte de la realidad, ya que el INML-CF únicamente registra los casos judicializados, además que un gran número de casos no son denunciados, en los últimos años han aumentado las cifras sobre delitos sexuales. El aumento de las denuncias puede deberse "a los cambios en la legislación y el clima social… [que]… concuerda con un interés creciente desde hace varios años, en la política pública para facilitar la denuncia y el acceso a los entes judiciales de estos casos" (*op. cit.*, p. 178).

Todas las fuentes de información oficial tienen una limitación al considerar únicamente los hechos delictuosos que ingresaron de forma efectiva al sistema penal y no la totalidad de los hechos.

La fracción de los hechos que no ingresa es lo que se denomina comúnmente subregistro o "cifra negra del delito". Su origen suele estar en la omisión de la denuncia, que puede darse por diversas razones. Entre otras, las que Sozzo señala son la creencia en que determinados delitos no justifican trámite administrativo, creencia en que la Policía o la Justicia son ineficientes o no van a solucionar el problema; creencia en que las fuerzas de seguridad locales pueden estar involucradas en el hecho; algún grado de involucramiento de la víctima en el hecho; temor por parte de esta a eventuales represalias o a atravesar situaciones de humillación o dolor; temor a ser estigmatizado o etiquetado (2000).

Las encuestas de victimización se convierten en una buena estrategia para mitigar el impacto del subregistro. Por medio de estas encuestas se pregunta al entrevistado si ha sido víctima de algún delito, él o alguno de los miembros de su familia conviviente, en una temporalidad establecida de acuerdo con el diseño de la investigación, y se indaga por las características del delito, si se hizo denuncia o no; en caso de omisión, las razones. Para el Informe Bogotá Cómo Vamos (2011) la Encuesta de Percepción Ciudadana, además, pregunta por la confianza en las distintas fuerzas de seguridad, percepción de seguridad y medidas de autoprotección adoptadas. Las encuestas de victimización, en general, no cubren todo el espectro de delitos, para el caso de Bogotá se indaga por los llamados delitos de mayor impacto: hurto a personas, robo a residencias, homicidio común, abuso sexual, venta de drogas, lesiones personales, violencia intrafamiliar, entre otros.

Informe de calidad

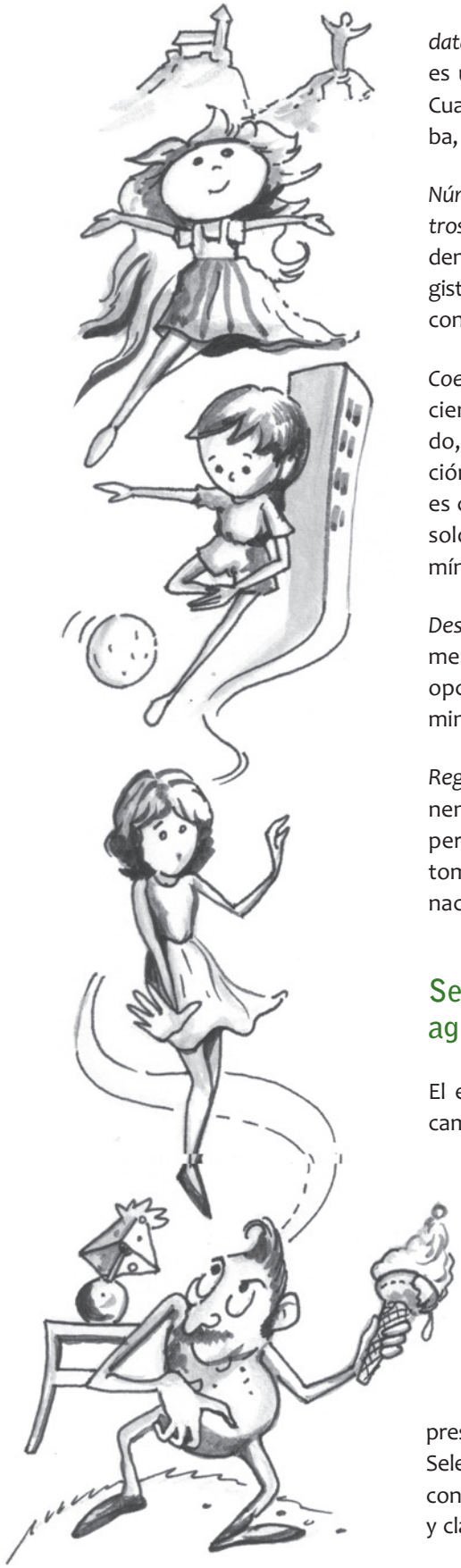
Se cargaron 50 campos y 4.425 registros al sistema de minería de datos. El informe de calidad da cuenta del porcentaje de registros con contenido informático. De los 50 campos, 12 presentan contenido deficiente, y en la mayoría de los casos por tener valores nulos son excluidos del análisis, aunque los campos con buen contenido informático pueden llegar a ser descalificados por tener una muy alta o muy baja variabilidad.

Además, se realizó una auditoría de los datos para determinar los campos que presentaran un nivel aceptable de contenido informacional y sus respectivos registros, para hacer el perfil de los determinantes estructurales del campo Presunto agresor. Para los campos seleccionados se determinan: los valores posibles que puede tomar cada campo; la descripción de cada uno de los valores; la frecuencia o cantidad de registros para cada valor posible; el nivel de completitud, es decir, la cantidad de registros vacíos, incompletos, erróneos, fuera de rango, y para los campos omitidos se explica la razón de su omisión.

Se aplica el algoritmo *Selección por características* para elegir las variables definidas como explicativas. El proceso de cribado implica eliminar predictores o casos que no aportan ninguna información útil teniendo en cuenta la relación predictor/objetivo. Las opciones de cribado se basan en atributos del campo en cuestión, sin contemplar la eficacia predictiva del campo objetivo seleccionado. Los campos cribados se excluyen de los cálculos utilizados para ordenar predictores por rangos y, opcionalmente, se pueden filtrar o eliminar de los datos utilizados en el modelado (apéndice 1). Los campos del archivo de datos de origen se cribaron en función de los siguientes criterios:

Porcentaje máximo de registros en una única categoría. Se cribaron los campos con demasiados registros dentro de la misma categoría en relación con el número total de registros. Si el 90% de los del

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia



dataset lleva a la misma categoría en determinado campo, no es útil incluir esta información para distinguir poblaciones. Cualquier campo que exceda el máximo especificado se criba, opción que solo se aplica para los campos categóricos.

Número máximo de categorías como un porcentaje de registros. Utilizado para los campos categóricos, aquellos con demasiadas categorías en relación con el número total de registros también fueron cribados. Si un 95% de las categorías contienen solo un único caso, el campo será de uso limitado.

Coeficiente mínimo de variación. Los campos con un coeficiente de varianza menor o igual que el mínimo especificado, son cribados. Esta medida es el cociente de la desviación típica del predictor dividida por su media. Si este valor es cercano a cero, no habrá mucha variabilidad. Utilizado solo a los campos de rango numérico, con un coeficiente mínimo de variación en 0,1.

Desviación típica mínima. Los campos con desviación típica menor o igual que el mínimo especificado se cribaron. Esta opción solo se aplica a campos de rango numérico, determinado por la desviación típica mínima en 0,09.

Registros con datos perdidos. Los registros o casos que tienen valores perdidos en el campo objetivo, o bien valores perdidos en todos los campos predictores, se excluyen automáticamente de todos los cálculos utilizados en la ordenación por rangos de los predictores.

Selección por características con "presunto agresor" como variable objetivo

El examen de contenido informacional se realizó con un campo objetivo o de salida. En minería de datos no se parte de la identificación de las variables dependientes y las variables independientes, todos y cada uno de los campos o atributos son susceptibles de ser explicados por los restantes. En este estudio se definió un objetivo para ser explicado por los demás campos.

Esta fase de cribado se limita a determinar cuáles campos son los potencialmente más explicativos del objetivo, que en este caso es el campo que registra el presunto agresor. Para este fin se aplican los algoritmos de Selección por características. El algoritmo filtra los campos con más de un porcentaje especificado de valores perdidos y clasifica los restantes según la importancia relativa para el

objetivo especificado (apéndice 3). No obstante, se seleccionaron unos campos que el modelo había descartado, por tener una categoría muy grande en el conjunto posible y demasiados valores perdidos, como estado civil y circunstancia, considerando que pueden llegar a ser predictores y por analizar la relación que guardan con la variable de salida.

Archivo de datos definitivo

El archivo de datos susceptible de análisis queda constituido por 4.425 registros originales y 15 campos, que aportan información de distintas dimensiones de la violencia sexual. Los campos que hacen referencia a la dimensión sociodemográfica contienen información sobre características de la víctima y su relación con el presunto agresor, representado por los atributos: edad, sexo, escolaridad, ocupación, identificación, estado civil, presunto agresor.

La dimensión circunstancial se constituye por información sobre el modo en que ocurrieron los hechos, considera atributos como escenario o lugar; circunstancia, es decir la situación inicial que originó o impulsó al agresor a cometer el hecho punible de acuerdo con los indicios, hallazgos o información suministrada por testigos en el lugar de los hechos; intervalo de hora y actividad que realizaba la víctima en el momento de la lesión. La dimensión de peritaje se compone de la información de dictámenes sexológicos proporcionados por el médico perito y considerados en el atributo posible delito sexual.

Técnicas utilizadas para el análisis de los datos

Agrupamiento o clustering

Con esta técnica se pretende generar unos conjuntos lo más homogéneos en su interior y heterogéneos entre sí. Se logra agrupando el conjunto de datos basándose en la similitud de los valores de sus atributos. Identifica regiones densamente pobladas, denominadas *clusters*, de acuerdo con alguna medida de distancia establecida (Chen, 1996) a la vez que maximiza la similitud de las instancias en cada *cluster* y minimiza la similitud entre *clusters* (Han & Kamber, 2001).

La técnica *clustering* ha sido estudiada en las áreas de la estadística (Jain & Dubes, 1988); *machine learning* (Fisher, 1996); base de datos espaciales y minería de datos (Ester et al., 1995; Cheeseman & Stutz, 1996). Entre los algoritmos de *clustering* más utilizados están *Self Organizing Maps* (SOM) o *Kohonen* y *K-Means*.

K-Means, es un método iterativo que busca formar *k clusters*, con *k* predeterminado antes del inicio del proceso. Comienza particionando los datos en *k* subconjuntos no vacíos, calcula el centroide de cada partición como el punto medio del *cluster* y asigna cada dato al *cluster* cuyo centroide sea el más próximo. Luego vuelve a particionar los datos iterativamente, hasta que no haya más datos que cambien de *cluster* de una iteración a la otra (Kaufman & Rousseeuw, 1990)[6]. *K-Means* es un método donde se construye una partición de una base de datos *D* de *n* objetos en un conjunto de *k* grupos, buscando optimizar el criterio de particionamiento elegido, en el cual cada grupo está representado por su centro (apéndice 2).

6 Otros algoritmos de *clustering* son *K-medoids* o PAM (*Partition Around Medoids*) y CLARA (*Clustering Large Applications*).

Algoritmo C5.0

Este método se diferencia por la forma en que realiza las pruebas sobre las variables. El algoritmo construye un árbol de decisión y evalúa la información de cada caso utilizando los criterios de entropía y ganancia o proporción de ganancia, según sea el caso.

Las formas en que realiza las pruebas a las variables pueden ser de tres tipos: i. La prueba 'estándar' para las variables discretas, con un resultado y una rama para cada valor posible de la variable. ii. Una prueba más compleja, basada en una variable discreta, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de para cada valor. iii. Si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados A <= Z y A > Z, para lo cual debe determinarse el valor límite Z (apéndice 3).

Resultados

Del archivo de datos obtenido a partir de *Selección por características*, se aplicó el algoritmo *K-Means* para agrupar 4.425 registros en 3 *clusters*, sobre los cuales se aplicó el algoritmo C5.0 para realizar una interpretación formal y definitiva.

El modelo *K-Medias* ofrece un método de análisis de *clusters* que permite conglomerar el conjunto de datos en distintos grupos cuando no se sabe *ex ante* cómo se comportan. A diferencia de la mayoría de los métodos de aprendizaje, estos modelos no utilizan un campo objetivo, aprendizaje no supervisado, que en lugar de intentar predecir un resultado busca revelar los patrones en el conjunto de campos de entrada. Los registros se agruparon con mayor similaridad en los valores intra-*cluster* y máxima disimililaridad entre valores inter-*cluster*[7]. En este estudio se realizaron 11 iteracciones consecutivas, hasta lograr la mejor adecuación a dichas similaridades y disimilaridades (apéndice 4).

Una vez se corre el modelo obtenemos los tres *clusters*. El *cluster-1* está conformado por 1.424 registros; el *cluster-2* por 1.560 registros y el *cluster-3* por 1.441. Están compuestos por diferente proporción de registros, pero con los mismos atributos: unidad local, edad, sexo, identificación, escolaridad, estado civil, intervalo de hora, escenario, actividad, circunstancia, dictamen topográfico, presunto agresor, número de agresores, posible delito sexual.

Los modelos de *cluster* se utilizan típicamente para buscar conjuntos de registros similares basados en los campos examinados. Los resultados pueden utilizarse para identificar las asociaciones implícitas que de manera directa no son observables y pasan inadvertidas. El resultado obtenido tras la ejecución de *K-Means* para 3 *clusters*, permite identificar que si bien la media del campo continuo para cada *cluster* se encuentra muy cerca de la media global, no ocurre lo mismo con las modas de los campos categóricos. Existe cierta alternancia entre las modas de los campos *actividad, edad, escolaridad, estado civil, intervalo, presunto agresor,* que parecen estar identificando los *clusters*.

Es de esperarse que con el número de campos involucrados el algoritmo no logre realizar un agrupamiento muy exhaustivo, ni llegue a elaborar *clusters* muy definidos. Sin embargo,

7 *K-Medias* empieza definiendo un conjunto de centros de conglomerados iniciales derivados de los datos. Después asigna cada registro al conglomerado de registros más similares, basándose en los valores de los campos de entrada de registros. Una vez asignados todos los casos, los centros de conglomerados se actualizan para reflejar el nuevo conjunto de registros asignados a cada conglomerado. Los registros se vuelven a comprobar para ver si se deben reasignar a otro conglomerado, y el proceso de iteración de conglomerado/asignación continúa, hasta que se alcanza el número máximo de iteraciones o el cambio entre una iteración y otra no sobrepasa el umbral especificado.

el examen contemplando todos los campos permite observar patrones a ese nivel de generalidad de un fenómeno multivariado y extremadamente dinámico, como lo es el delito en general y la violencia sexual en particular.

El índice de importación señala que cuanto mayor sea la medida de importancia, menor probabilidad habrá de que la variación de un campo entre *cluster* se presente como producto de la probabilidad, y mayor probabilidad de que exista una diferencia subyacente. Por esta razón el análisis se concentra en los campos con un mayor nivel de variación entre los *clusters*[8]. Como los centroides no necesariamente representan la combinación de atributos más frecuentes, se hace necesaria una aproximación más detallada para caracterizarlos, que debe empezar por asignar una etiqueta o nombre que sugieran dichas características. Los *clusters* configurados se han llamado: *Cluster* 1: *violencia sexual en situación de incesto*; *Cluster* 2: *violencia sexual en situación anonimia*; *Cluster* 3: *violencia sexual en situación de estructura familiar*.

Cluster 1: violencia sexual en situación de incesto

Recuento: 1.424; porcentaje del total: 47,45%

Tabla 1. Cluster: violencia sexual en situación de incesto

Campo	Centroide
Edad	3 a 12 años
Nivel de escolaridad	Primaria
Intervalo de hora	12:01 a 18:00
Actividad	Hogar - vital
Escenario	Vivienda
Presunto agresor	Familia nuclear
Posible delito sexual	Probable abuso sexual

La víctima de violencia sexual en situación de incesto se caracteriza porque se encuentra entre los 3 y 12 años de edad; cuenta con nivel de escolaridad en primaria y se encuentra en la vivienda entre las 12:01 y las 18:00 horas; realizando actividades vitales; el presunto agresor está dentro de la familia nuclear; en consecuencia, es probable abuso sexual (apéndice 5).

Cluster 2: violencia sexual en situación de anonimia

Recuento: 1.560; porcentaje del total: 54,45%

Tabla 2. Cluster: violencia sexual en situación de anonimia

Campo	Centroide
Edad	13 a 15 años; 16 a 24 años
Nivel de escolaridad	Secundaria
Intervalo de hora	18:01 a 24:00

8 La importancia se calcula como 1 menos el valor *p*, donde el valor de probabilidad se obtiene a partir de las pruebas *T* (para campos de rango) y las chi-cuadrado (para campos discretos).

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

Campo	Centroide
Actividad	Vital - tiempo libre
Escenario	Espacio público
Presunto agresor	Desconocido
Posible delito sexual	Probable asalto sexual

Es el *cluster* que más registros agrupa y el más parecido a la media global. Está caracterizado porque la víctima tiene entre 13 a 15 años, y entre los 16 a 24 años; la escolaridad es de nivel secundaria; el hecho ocurre mientras realizan actividades vitales y de tiempo libre; durante las 18:01 y las 24:00 horas; en el espacio público; el agresor es desconocido o el ex (ex compañero, ex novio, ex esposo, ex amante); considerándose un probable asalto sexual (apéndice 6).

Cluster 3: violencia sexual en situación de estructura familiar

Recuento: 1.424; porcentaje del total: 47,45%

Tabla 3. Cluster: Violencia sexual en situación de estructura familiar

Campo	Centroide
Edad	3 a 12 años
Nivel de escolaridad	Primaria incompleta
Actividad	Actividad vital
Escenario	Vivienda
Presunto agresor	Familia nuclear; familiares políticos, consanguíneos

Es el más difuso de los *clusters*, ya que la mayoría de sus registros presentan traslape con los de las otras agrupaciones. Sin embargo, se intentó aislar por ciertos perfiles generales. La víctima tiene entre 3 a 12 años de edad; con primaria incompleta; el hecho ocurre mientras realiza actividades vitales; mientras está en la vivienda, y el presunto agresor está dentro del núcleo familiar u otros familiares políticos o consanguíneos (apéndice 7).

Aplicación del algoritmo C5.0: árbol de decisión

Utilizamos el algoritmo C5.0 para preseleccionar los atributos que serán utilizados en algoritmos *Top Down Induction of Decision Trees* (TDIDT). Este algoritmo califica de manera eficiente los registros de los campos involucrados, generando un árbol de decisión, que tiene una profundidad de nivel 7 con 263 hojas. La utilización de este algoritmo confirma que los *clusters* determinados por *K-Means* responden a un criterio subyacente en los datos y no son producto de factores aleatorios. Además, el 10% de las reglas de clasificación extraídas del árbol clasifican el 70% de las instancias (4.425). El nodo principal es el atributo que da cuenta del *presunto agresor*. La raíz que le corresponde se dirige hacia los nodos de *circunstancia*. Del nodo *conflictividad* se desprende la rama *escenario* que se dirige a los nodos *centros educativos*, *lugar de comercio o bebidas*, *lugar público*, *lugar de hospedaje*, *vehículo y vivienda*.

Las reglas fueron contrastadas con publicaciones anteriores y permitieron la confirmación de las interpretaciones realizadas hasta el momento. Se suele clasificar la violencia sexual en dos grupos, de acuerdo con el vínculo existente entre la víctima y el presunto agresor: casos en los que víctima y agresor se conocen; casos en los que no se conocen víctima y agresor (apéndice 8).

El primer grupo está representado por el *cluster* 2, mientras que el segundo está contenido por los *clusters* 3 y 1. Las diferencias entre estos últimos indican cierta relación entre el lugar de ocurrencia del hecho y el presunto agresor. Si este es desconocido, es más probable que el escenario sea un espacio público. Si el agresor es del núcleo familiar, entonces hay mayor posibilidad de que sea en la vivienda; aunque si el agresor es conocido con algún nivel de afectividad o con quien se tuvo alguna relación afectiva, entonces la distinción se hace borrosa, pero se separa a un nivel de exhaustividad más exigente.

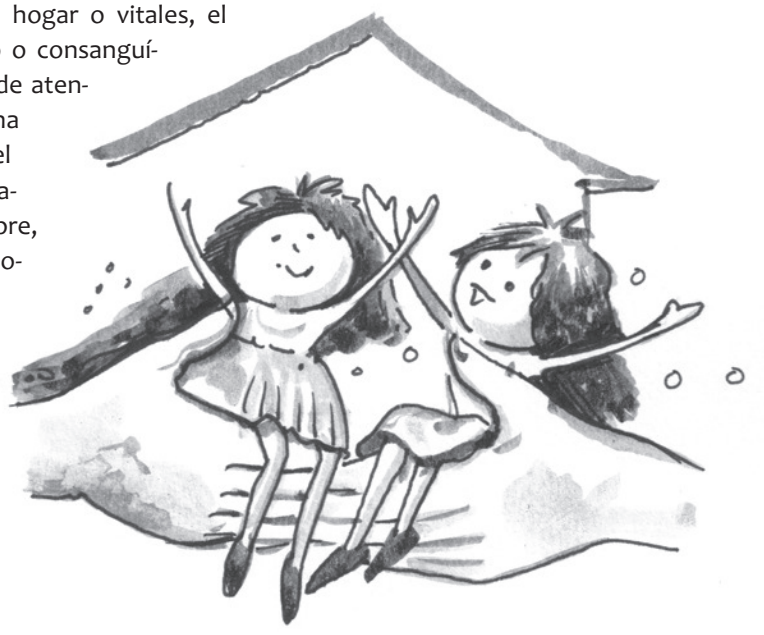
La conclusión, de acuerdo con la literatura de investigaciones especializadas, es que se trata de dos tipos de violencias sexuales: una en situación de incesto y otra en estado de anonimia. En medio de estos dos grupos están los casos difíciles de asignar *a priori* a una u otra modalidad sin la ayuda de un algoritmo clasificatorio que permita identificarlo como violencia sexual en situación de estructura familiar.

Características sociodemográficas y contextuales

Si bien no fue posible desarrollar todas las líneas de análisis, la consulta de las reglas generadas por el modelo y el árbol de clasificación puede ser recurrente y abierta. Es pertinente focalizar el análisis en tres poblaciones diferenciadas, con referencia al presunto agresor, considerando variables sociodemográficas y de contexto, perfiladas en tres agrupamientos.

Una población eminentemente femenina, característica que aplica para el 84% del total de los casos de estudio, entre los 3 y 12 años de edad, con estudios de primaria, cuyo presunto agresor pertenece al núcleo familiar. La población que se encuentra entre los 13 y 15 años, tiene estudios secundarios, el presunto agresor es desconocido u otros familiares políticos o consanguíneos. Por último una población entre los 16 y 24 años, con secundaria incompleta y con ocupación como estudiante o ama de casa, cuyo presunto agresor es conocido con algún nivel de afectividad, o con quien ha tenido una relación afectiva, aunque durante la ocurrencia del hecho esta no se mantiene (apéndice 9).

Se encuentra una población que es víctima en la vivienda, mientras realiza actividades del hogar o vitales, el agresor es otro familiar político o consanguíneo y es atendida en la unidad de atención al menor. Se presenta una población que es víctima en el espacio público, durante la realización de actividades de tiempo libre, y el agresor es dudoso o desconocido. Por último, la población que es víctima en lugares de hospedaje, vehículo y/o lugar público, mientras realiza actividades de tiempo libre, cuyo agresor tiene o tuvo alguna relación con algún nivel de afectividad con la víctima (apéndice 10).



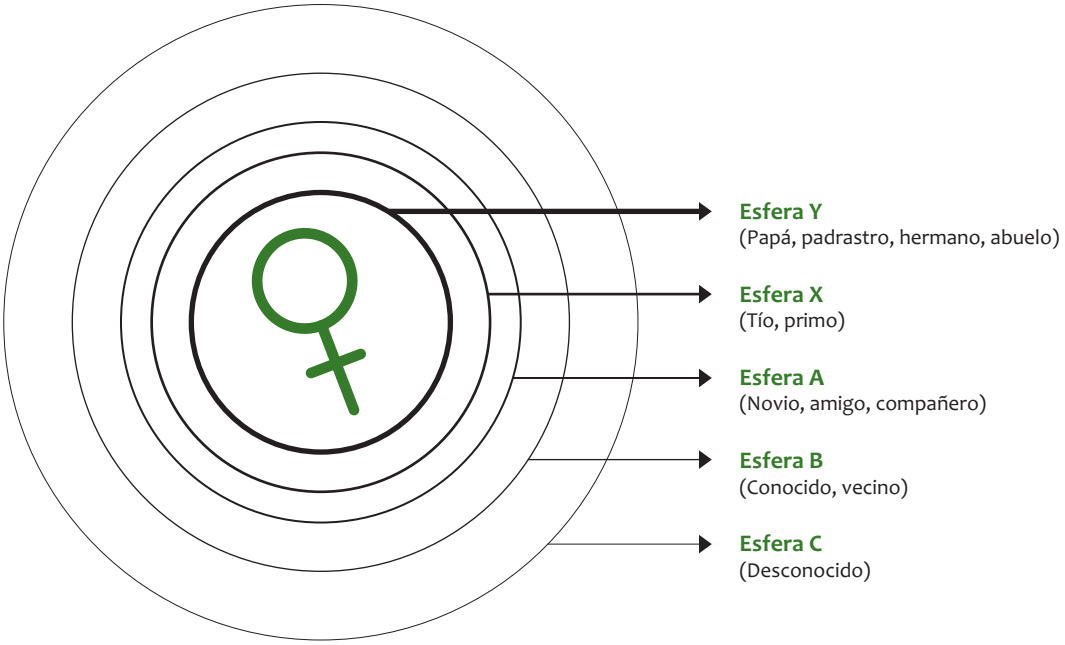
ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

Hallazgos

Las técnicas aplicadas a los datos permitieron realizar un análisis sobre los comportamientos y las transformaciones del fenómeno, teniendo en cuenta similaridades y disimilaridades. Se derivan tres patrones de conducta del fenómeno de la violencia sexual: *violencia sexual en situación de incesto, violencia sexual en situación de anonimia* y *violencia sexual en situación de estructura familiar*.

La *violencia sexual en situación de incesto* ocurre, básicamente, de puertas para dentro y el perpetrador comete al mismo tiempo *incesto*. La *violencia sexual en situación de anonimia*, en lo fundamental, se produce de puertas para fuera, en espacios públicos o "terceros lugares" y el perpetrador se apoya en la anonimia imperante en el espacio público de la ciudad. La *violencia sexual en situación de estructura familiar* es cometida por una persona que forma parte de la familia ampliada y tiene un parentesco político o consanguíneo con la víctima. También entra en esta categoría el delito cometido por una persona que tuvo algún vínculo afectivo-sentimental con la víctima pero que en el momento del hecho ya no existe dicho vínculo.

Gráfico 1. Esferas de proximidad con la víctima



La institución de la familia está siendo sometida a transformaciones que llevan a considerar las nociones clásicas de parentesco como poco ajustadas a la realidad empírica de la estructura familiar. La dinámica de esta institución social muestra que la separación entre pariente consanguíneo o civil no es adecuada, y obliga a buscar otro marco analítico para comprender los aspectos de la estructura social relacionados con la violencia sexual. Si bien no se intenta discutir o definir un nuevo modelo analítico para la estructura familiar, las técnicas aplicadas muestran otro criterio de clasificación, a saber: los niveles o esferas de proximidad con la víctima en el centro de dichas esferas (gráfico 1).

Esta clasificación empírica corta al través la clasificación tradicional basada en el parentesco. Si, como se aprecia, la tendencia en la estructura familiar, en particular, y la estructura social, en general, se dirige hacia mayores niveles de individuación, es claro que estos solo son posibles

gracias al incremento y fortalecimiento de *instituciones de propósito*[9]. Estas deben penetrar con mayor fuerza y de manera más efectiva los núcleos familiares, y la composición de los hogares para inhibir la propensión a la violencia sexual en su expresión *en situación de incesto*. Apelar al reforzamiento de los lazos familiares tradicionales es algo un tanto utópico, con la emergencia marcada de la individuación en los tiempos actuales.

La *violencia sexual en situación de anonimia* se apoya en el factor ecológico. Las personas para realizar sus actividades deben desplazarse por sectores distantes entre sí e inmersos en el espacio urbano, debido a este factor, tarde o temprano, quedan sumergidas en la anonimia. Sería pertinente articular la estructura institucional de forma que las personas puedan configurar sus trayectorias, en especial el movimiento pendular diario de residencia-trabajo o estudio, para que en su desplazamiento por el espacio urbano cuenten con puntos focales institucionales que brinden respaldo.

La *violencia sexual en situación de estructura familiar* demanda el control social de manera descentralizada por medio de normas sociales. La regulación informal, por medio de estas, puede llevar a configurar la situación de una manera que refuerce los aspectos sociales y culturales que inhiben la conducta desviada. Focalizando la intervención hacia la prevención, con campañas de muy bajo costo, se puede generar o reforzar normas sociales que regulan el trato con las personas con las que se ha tenido una relación con algún nivel de afectividad pero que por distintas circunstancias esta ya no existe.

Discusión

La aplicación de la metodología de minería de datos a la información existente y registrada sobre delitos en la ciudad de Bogotá, resalta el valor agregado de este tipo de análisis para la comprensión y generación de nuevo conocimiento sobre el fenómeno delictivo. Los resultados experimentales obtenidos han sido contrastados con la investigación y reflexiones de especialistas del INML-CF y el CEACSC, que ha permitido confirmar conceptos preexistentes y generar nuevas piezas o trozos de conocimiento. Al respecto se han identificado tres patrones de violencia sexual con base en los hechos registrados por el INML-CF en Bogotá durante el año 2007. Los patrones de *violencia sexual en situación de incesto* y *la violencia sexual en situación de anonimia* validan claramente el conocimiento profesional a la fecha que señala una clasificación entre abuso sexual y asalto sexual. La *violencia sexual en situación de estructura familiar* señala un patrón en los datos que, si bien es borroso, aporta nuevo conocimiento en tanto pone atención a la trasformación de la estructura familiar y de parentesco, y las implicaciones que esta transformación acarrea para la comisión de actos punibles, como el delito sexual.

Según la Organización Mundial de la Salud, la violencia es una de las principales causas de muerte en la población entre los 15 y 44 años de edad. La violencia ocurre en los hogares, en lugares de trabajo e incluso en instituciones médicas y sociales, lo cual puede generar en alguna medida que las víctimas se vean obligadas, por convenciones o presiones sociales, a guardar silencio. El bienestar social

9 Instituciones de propósito específico son aquellas que emergen a partir del socavamiento de la autoridad con la que, otrora, contaban las instituciones primarias, como la familia. Esta institución primaria ha venido perdiendo autoridad en tanto emergen instituciones de propósito específico, como la institución de la educación, el trabajo, el entretenimiento, que cumplen funciones que antes ejercía la primera (Coleman, 1990).

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

de las víctimas puede verse afectado por la estigmatización y aislamiento por parte de sus familias u otras personas de su círculo social (*op. cit.*, p. 13). Esta situación puede aumentar la violencia y generar altos niveles de impunidad, por la ausencia de información y el subregistro de denuncias. Para obtener mayor claridad en la configuración del *cluster 3* identificado, en particular, una estrategia significativa puede ser complementar los registros de entidades oficiales con datos de encuestas de victimización. Esta estrategia en estudios más generales ha sido representativa en investigaciones sobre violencia, delito y conflictividad (Dammert, 2010).

Si bien el aumento en los registros de los casos de violencia sexual es favorable a las sobrestimaciones, por el incremento del 65,9% al comparar las tasas de 1997 y de 2007 (González, J., 2007), se insiste en la importancia del trabajo articulado para la formulación de respuestas integrales y orientación de política pública, basadas en una revisión teórica, apoyada en el metaanálisis y sustentada en investigaciones, para generar conocimiento en la comprensión de la realidad del hecho violento.

La utilización de la minería de datos para el análisis de información criminal ha demostrado ser prometedora, teniendo en cuenta que sus distintas aplicaciones han permitido relacionar delitos de autoría desconocida según el *modus operandi*, optimizar la locación de los recursos policiales y detectar grupos delictivos organizados. Es importante que la postura oficial sea concebir la información criminal no solo como "termómetro" de la inseguridad sino también como herramienta fundamental para la toma de decisiones. Cuanto mejor sea la calidad de la información, más acertadas serán las decisiones y más efectivamente se podrá reducir los niveles de violencia.

Estas nuevas herramientas de minería de datos ofrecen un soporte en las relaciones y modelos, e interrelaciones de las variables y aspectos que no se observan o con frecuencia se han mantenido aislados. Lo cual puede orientar el proceso de toma de decisiones sobre la base del estudio de variables que por lo general se mantienen separadas. Estas herramientas permiten, en grandes volúmenes de datos, realizar el ejercicio, tanto deseable como difícil de lograr, en el análisis sociológico de los fenómenos: separar lo que comúnmente se une, unir lo que comúnmente se separa.

Bibliografía

Centro de Estudio y Análisis en Convivencia y Seguridad Ciudadana (2008). *Violencia intrafamiliar y abuso sexual*. No publicado.

Centro Internacional para la Prevención de la Criminalidad (2008). *Informe internacional Prevención de la criminalidad y seguridad cotidiana: tendencias y perspectivas.* Centro Internacional para la Prevención de la Criminalidad. Canadá: CIPC.

Cheeseman, P. & Stutz, J. Bayesian classification (AutoClass): Theory and results. En: Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). *Advances in Knowledge Discovery and Data Mining* (pp. 153-180). U.S.A.: American Association for Artificial Intelligence Menlo Park.

Chen, M. S., Han, J. & Yu, P. (1996). Data mining: An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8 (6): 866-883.

Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., & Chau, M. (2004). Crime Data Mining: A General Framework and Some Examples*. IEEE Computer Society*, 37 (4): 50-56.

Congreso de Colombia (2000). Ley 599 de 2000. Código Penal. Artículos 138-141; 205-210 [versión electrónica]. Recuperado el 19 de diciembre del 2011 de: http://alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=6388.

Coleman, J. (1990). *Foundations of Social Theory.* Cambridge, Mass.: Harvard University Press.

Dammert, L., Salazar, F., Montt, C. & González, P. A. (2010). *Crimen e inseguridad: indicadores para las Américas*. Santiago, Chile: Flacso-Chile/Banco Interamericano de Desarrollo.

Ester, M., Kriegel, H. P. & Xu, X. (1995). Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. En: *Proc. 4th Int. Symp. on Large Spatial Databases (SSD'95)* (pp. 67-82). Portland, Maine, EE. UU.

Fisher, D. (1996). *Iterative optimization and simplification of hierarchical clusterings*. Departament of Computer Science. EE. UU.: Vanderbilt University, Nashville.

Han, J. & Kamber, M. (2001). *Data mining: Concepts and techniques*. United States of America: Morgan Kauffmann Publishers.

González, J. (2007). Informes periciales por presunto delito sexual. *Forensis. Datos para la* vida, pp. 143-178.

Jain, A. & Dubes, R. (1988). *Algorithms for Clustering Data*. USA: Prentice Hall.

Kantardzic, M. (2002). *Data Mining: Concepts, models, methods and algorithms.* IEEE Press & John Wiley.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis*. New York: Wiley-Interscience.

Ng, R. T. & Han, J. (1994). Efficient and effective clustering method for spatial data mining. En: Proc. *Int. Conf. Very Large Data Bases*, pp. 144/155. Santiago de Chile, Chile.

Ochoa, M. A. (2004). *Herramientas inteligentes para la explotación de información*. Trabajo final: Especialidad en Ingeniería en Sistemas Expertos, Instituto Tecnológico de Buenos Aires (ITBA).

Organización Panamericana de la Salud (2003). *Informe mundial sobre la violencia y la salud*. Washington, D.C.: Oficina Regional para las Américas de la Organización Mundial de la Salud.

Pinzón, D. (2009). La violencia de género y la violencia sexual en el conflicto armado colombiano: indagando sobre sus manifestaciones. En: J. Restrepo & D. Aponte (Eds.). *Guerra y violencias en Colombia*. Bogotá: Pontifica Universidad Javeriana.

Sozzo, M. (2000). *Pintando a través de números: fuentes estadísticas de conocimiento y gobierno democrático de la cuestión criminal en Argentina*. Recuperado en septiembre del 2009 de : http://www.ilsed.org/index.php?option=com_docman&task=doc_view&gid=159&itemid=44.

Sozzo, M. (Director) (2005). *Policía, violencia, democracia: ensayos sociológicos*. 1a ed. Santa Fe: Universidad Nacional del Litoral.

Tan, P. N., Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining*. USA: Addison-Wesley.

Zeleznikow, J. (2005). *Using Data Mining to Detect Criminal Networks.* Recuperado en septiembre del 2009 de: http://www.aic.gov.au/conferences/occasional/2005-04.zeleznikow.html.

Zhang, T., Ramakrishnan, R. & Livny, M. (1996). *BIRCH: an efficient data clustering method for very large databases*. En Proc. ACM-SIGMOD Int. Conf. Management of Data. Montreal, Canadá.

APÉNDICES

Apéndice 1. Campos seleccionados como predictores por el algoritmo de Selección por características con presunto agresor como objetivo

		Rango	Campo	Tipo	Importancia	Valor
1	true	1	OCUPACIÓN2	set	Importante	1,0
2	true	2	ESCENARIO1	set	Importante	1,0
3	true	3	DTO_TOPOGR1	set	Importante	1,0
4	true	4	ACTIVIDAD1	set	Importante	1,0
5	true	5	ULOCAL	set	Importante	1,0
6	true	6	ESCOLARIDAD	set	Importante	1,0
7	true	7	POSDELSEX	set	Importante	1,0
8	true	8	SEXO	flag	Importante	1,0
9	true	9	EDAD1	range	Importante	1,0
10	true	10	PMV	range	Importante	1,0
11	true	11	CARISEX	range	Importante	1,0
12	true	12	GRUPOVULNE	range	Importante	1,0
13	true	13	CONSENT	range	Importante	1,0
14	true	14	N_AGRESORE1	range	Importante	1,0
15	true	15	INTERVALO1	set	Importante	1,0
16	false	16	ATENCIÓN_M	range	Sin importancia	0,835
17	false	17	IDENTIFICACIÓN	flag	Sin importancia	0,832
18	false	18	SEXORAL	range	Sin importancia	0,794
19	false	19	POMV	range	Sin importancia	0,546
20	false	20	ACTOSP	range	Sin importancia	0,307

Apéndice 2: Algoritmo *K-Means*

El objetivo que se intenta alcanzar es minimizar la varianza total intra-grupo o la función de error cuadrático cuya notación sucinta es:

$$V = \sum_{i=0}^{k} \sum_{j \epsilon S_i} \left| x_j - \mu_i \right|^2$$

Donde existen *k* grupos *Si*, i = 1,2,..., *k* y *μi* son el punto medio o centroide de todos los puntos, *Xj* ϵ *Si*. *K-Means* comienza particionando los datos en *k* subconjuntos no vacíos, aleatoriamente o usando alguna heurística. Luego calcula el centroide de cada partición como el punto medio del *cluster* y asigna cada dato al *cluster* cuyo centroide sea el más próximo. Luego los centroides son recalculados para los grupos nuevos y el algoritmo se repite hasta la convergencia, la cual es obtenida cuando no haya más datos que cambien de grupo de una iteración a otra.

Para calcular el centroide más cercano a cada punto se debe utilizar una función de distancia. Para datos reales se suele utilizar la distancia euclídea. Para datos categóricos se debe establecer una función específica de distancia para ese conjunto de datos. Algunas de las opciones puede ser utilizar una matriz de distancias predefinidas o una función heurística.

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

El algoritmo no garantiza que se obtenga un óptimo global. La calidad de la solución final depende principalmente del conjunto inicial de grupos. Debido a esto, se suelen realizar varias ejecuciones del algoritmo con distintos conjuntos iniciales, que permita obtener una mejor solución.

Dado *k*, el algoritmo *K-Means* se implementa en 4 pasos (Tan, Steinbach & Kumar, 2005)[10]:

i. Particionar los objetos en *k* subconjuntos no vacíos.
ii. Computar los centroides de los *clusters* de la partición corriente. El centroide es el centro (punto medio) del *cluster*.
iii. Asignar cada objeto al *cluster* cuyo centroide sea más cercano.
iv. Volver al paso 2, parar cuando no haya más reasignaciones.

Apéndice 3: Algoritmo C5.0

Pseudocódigo del algoritmo C5.0

Las características particulares de este método, que lo diferencian de su antecesor, son:

El algoritmo del método C5.0 para la construcción de árboles de decisión, a grandes rasgos es muy similar al del ID3. Varía en la manera en que realiza las pruebas sobre las variables.

Función C5.0

(R: conjunto de atributos no clasificadores, C: atributo clasificador, S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
Si S está vacío,
Devolver un único nodo con Valor Falla;
Si todos los registros de S tienen el mismo valor para el atributo clasificador,
Devolver un único nodo con dicho valor;
Si R está vacío,
Devolver un único nodo con el valor más frecuente del atributo clasificador en los registros de S
[Nota: habrá errores, es decir, registros que no estarán bien clasificados en este caso];
Si R no está vacío,
D ← atributo con mayor Proporción de Ganancia (D,S) entre los atributos de R;
Sean {dj | j = 1,2,...., m} los valores del atributo D;
Sean {dj | j = 1,2,...., m} los subconjuntos de S correspondientes a los valores de dj respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d1,
d2,...,dm, que van respectivamente a los árboles
C4.5(R-{D}, C, Sl), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fin

10 *K-Means* es ampliamente utilizado en la explotación de datos, en la cuantificación de vectores, para cuantificar variables reales en *k* rangos no uniformes y para reducir el número de colores en una imagen.

Apéndice 4: Resumen del modelo de *K-Medias*

Análisis

Número de conglomerados: 3

Iteración	Error
1	1,812
2	0,336
3	0,184
4	0,09
5	0,14
6	0,193
7	0,124
8	0,026
9	0,006
10	0,001
11	0,0

Campos

Entradas

ULOCAL
EDAD1
SEXO
IDENTIFICACIÓN
ESCOLARIDAD
ESTADO_CIV
INTERVALO1
ESCENARIO
ACTIVIDAD
CIRCUNSTAN
DTO_TOPOGR
P_AGRESOR
N_AGRESORE
POSDELSEX

Configuración de creación

Utilizar los datos en particiones: falso
Número de conglomerados especificado: 3
Generar campo de distancia: verdadero
Mostrar proximidad de conglomerados: verdadero
Etiqueta de conglomerado: cadena
Prefijo de etiqueta: *cluster*
Optimizar: memoria
Modo: simple

Resumen de entrenamiento

Tipo de modelo: K-Medias

Ruta: Ruta1

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

Apéndice 5: Fragmento del Algoritmo *K-Medias - cluster 1*

1.424 registros

- N_AGRESORE
  - Media = 0,739
  - Desviación típica = 0,507
- SEXO
  - FEMENINO (74,93%)
  - FEMENINO 74,93%
  - MASCULINO 25,07%
- ACTIVIDAD
  - OTRA ACTIVIDAD (44,31%)
  - ACTIVIDAD VITAL 16,36%
  - APRENDIZAJE 3,65%
  - DELICTIVA 0,07%
  - DEPORTIVA 0,07%
  - HOGAR 13,97%
  - MISIÓN HUMANITARIA 0,07%
  - OTRA ACTIVIDAD 44,31%
  - QUEHACER NO REMUNERADO 0,14%
  - RETENCIÓN ILEGAL 0%
  - SIN INFORMACIÓN 15,31%
  - TIEMPO LIBRE 6,04%
  - TRABAJA 0%
  - TRANSPORTE AL TRABAJO 0%
- CIRCUNSTAN
  - OTROS (73,3%)
  - ATRACO CALLEJERO 0,42%
  - CONFLICTIVIDAD 1,55%
  - ENFRENTAMIENTO ARMADO 0,42%
  - HURTO 0,21%
  - OTROS 73,3%
  - SECUESTRO 0,07%
  - SIN INFORMACIÓN 24,03%
- DTO_TOPOGR
  - SIN LESIONES (89,12%)
  - ÁREA GENITAL/PARAGENITAL 6,46%
  - POLITRAUMA 1,19%
  - SIN LESIONES 89,12%
  - TRAUMA ABDOMEN 0,14%
  - TRAUMA ÁREA PÉLVICA 0,7%
  - TRAUMA CRANEANO 0,14%
  - TRAUMA CUELLO 0,14%
  - TRAUMA FACIAL 0,49%
  - TRAUMA MIEMBROS 0,98%
  - TRAUMA TÓRAX 0,63%

Apéndice 6: Fragmento del Algoritmo *K-Medias - cluster 2*

1.560 registros

- N_AGRESORE
  - Media = 0,813
  - Desviación típica = 0,768
- SEXO
  - FEMENINO (93,01%)
  - FEMENINO 93,01%
  - MASCULINO 6,99%
- ACTIVIDAD
  - OTRA ACTIVIDAD (51,28%)
  - ACTIVIDAD VITAL 12,88%
  - APRENDIZAJE 1,6%
  - DELICTIVA 0,13%
  - DEPORTIVA 0,06%
  - HOGAR 4,04%
  - MISIÓN HUMANITARIA 0%
  - OTRA ACTIVIDAD 51,28%
  - QUEHACER NO REMUNERADO 0,64%
  - RETENCIÓN ILEGAL 0,13%
  - SIN INFORMACIÓN 16,99%
  - TIEMPO LIBRE 10,58%
  - TRABAJA 0,38%
  - TRANSPORTE AL TRABAJO 1,28%
- CIRCUNSTAN
  - OTROS (58,79%)
  - ATRACO CALLEJERO 8,22%
  - CONFLICTIVIDAD 3,34%
  - ENFRENTAMIENTO ARMADO 1,8%
  - HURTO 0,83%
  - OTROS 58,79%
  - SECUESTRO 1,86%
  - SIN INFORMACIÓN 25,16%
- DTO_TOPOGR
  - SIN LESIONES (65,73%)
  - ÁREA GENITAL/PARAGENITAL 10,14%
  - POLITRAUMA 7,83%
  - SIN LESIONES 65,73%
  - TRAUMA ABDOMEN 0,77%
  - TRAUMA ÁREA PÉLVICA 1,54%
  - TRAUMA CRANEANO 0,19%
  - TRAUMA CUELLO 1,41%
  - TRAUMA FACIAL 3,79%
  - TRAUMA MIEMBROS 6,42%
  - TRAUMA TÓRAX 2,18%
- EDAD1
  - 16 A 24 (42,31%)
  - 0 A 2 1,03%
  - 3 A 12 5,19%
  - 13 A 15 36,35%

ISSN 1794-3108. Rev. crim., Volumen 53, número 2, julio-diciembre 2011, Bogotá, D. C., Colombia

Apéndice 7: Fragmento del Algoritmo K-Medias - *cluster 3*

1.441 registros

- N_AGRESORE
  - Media = 0,578
  - Desviación típica = 0,547
- SEXO
  - FEMENINO (82,72%)
  - FEMENINO 82,72%
  - MASCULINO 17,28%
- ACTIVIDAD
  - ACTIVIDAD VITAL (55,03%)
  - ACTIVIDAD VITAL 55,03%
  - APRENDIZAJE 0,07%
  - DELICTIVA 0%
  - DEPORTIVA 0,07%
  - HOGAR 0,49%
  - MISIÓN HUMANITARIA 0%
  - OTRA ACTIVIDAD 11,38%
  - QUEHACER NO REMUNERADO 0%
  - RETENCIÓN ILEGAL 0%
  - SIN INFORMACIÓN 32,82%
  - TIEMPO LIBRE 0,14%
  - TRABAJA 0%
  - TRANSPORTE AL TRABAJO 0%
- CIRCUNSTAN
  - OTROS (90,49%)
  - ATRACO CALLEJERO 0,42%
  - CONFLICTIVIDAD 0,14%
  - ENFRENTAMIENTO ARMADO 0,07%
  - HURTO 0%
  - OTROS 90,49%
  - SECUESTRO 0%
  - SIN INFORMACIÓN 8,88%
- DTO_TOPOGR
  - SIN LESIONES (62,57%)
  - ÁREA GENITAL/PARAGENITAL 35,35%
  - POLITRAUMA 0,07%
  - SIN LESIONES 62,57%
  - TRAUMA ABDOMEN 0%
  - TRAUMA ÁREA PÉLVICA 0,9%
  - TRAUMA CRANEANO 0%
  - TRAUMA CUELLO 0,07%
  - TRAUMA FACIAL 0,21%
  - TRAUMA MIEMBROS 0,69%
  - TRAUMA TÓRAX 0,14%
- EDAD1
  - 3 A 12 (72,17%)
  - 0 A 2 4,23%
  - 3 A 12 72,17%
  - 13 A 15 15,75%

Apéndice 8: Fragmento reglas modelo C5.0 P_AGRESOR

Reglas para CONOCIDO CON ALGÚN TRATO - contiene 28 regla(s)

Regla 1 para CONOCIDO CON ALGÚN TRATO

si ESCENARIO = CENTROS EDUCATIVOS y CIRCUNSTAN = CONFLICTIVIDAD entonces CONOCIDO CON ALGÚN TRATO

Regla 2 para CONOCIDO CON ALGÚN TRATO

si ULOCAL = ATENCIÓN AL MENOR y EDAD1 = 3 A 12 e INTERVALO1 = 12:01 a las 18:00 y ACTIVIDAD = OTRA ACTIVIDAD y DTO_TOPOGR = ÁREA GENITAL/PARAGENITAL entonces CONOCIDO CON ALGÚN TRATO

Regla 3 para CONOCIDO CON ALGÚN TRATO

si ULOCAL = DELITOS SEXUALES y SEXO = MASCULINO e INTERVALO1 = 06:01 a las 12:00 entonces CONOCIDO CON ALGÚN TRATO

Regla 4 para CONOCIDO CON ALGÚN TRATO

si ULOCAL = URI TOBERÍN - USAQUÉN y EDAD1 = 3 A 12 e INTERVALO1 = 18:01 a las 24:00 y ESCENARIO = VIVIENDA y CIRCUNSTAN = OTROS entonces CONOCIDO CON ALGÚN TRATO

Regla 5 para CONOCIDO CON ALGÚN TRATO

si ESCOLARIDAD = SECUNDARIA INCOMPLETA y OCUPACIÓN = EMPLEADO (A) y ESCENARIO = VIVIENDA y CIRCUNSTAN = OTROS y DTO_TOPOGR = SIN LESIONES y POSDELSEX = PROBABLE ASALTO SEXUAL entonces CONOCIDO CON ALGÚN TRATO

Regla 6 para CONOCIDO CON ALGÚN TRATO

si EDAD1 = 13 A 15 y SEXO = MASCULINO y ESCENARIO = SIN INFORMACIÓN y ACTIVIDAD = OTRA ACTIVIDAD entonces CONOCIDO CON ALGÚN TRATO

Regla 7 para CONOCIDO CON ALGÚN TRATO

si ULOCAL = DELITOS SEXUALES y OCUPACIÓN = ESTUDIANTE SECUNDARIA e INTERVALO1 = 12:01 a las 18:00 y ESCENARIO = VIVIENDA y ACTIVIDAD = OTRA ACTIVIDAD y DTO_TOPOGR = SIN LESIONES entonces CONOCIDO CON ALGÚN TRATO

Regla 8 para CONOCIDO CON ALGÚN TRATO

si ULOCAL = CJ CIUDAD BOLÍVAR e INTERVALO1 = SIN DATO y POSDELSEX = PROBABLE ASALTO SEXUAL entonces CONOCIDO CON ALGÚN TRATO

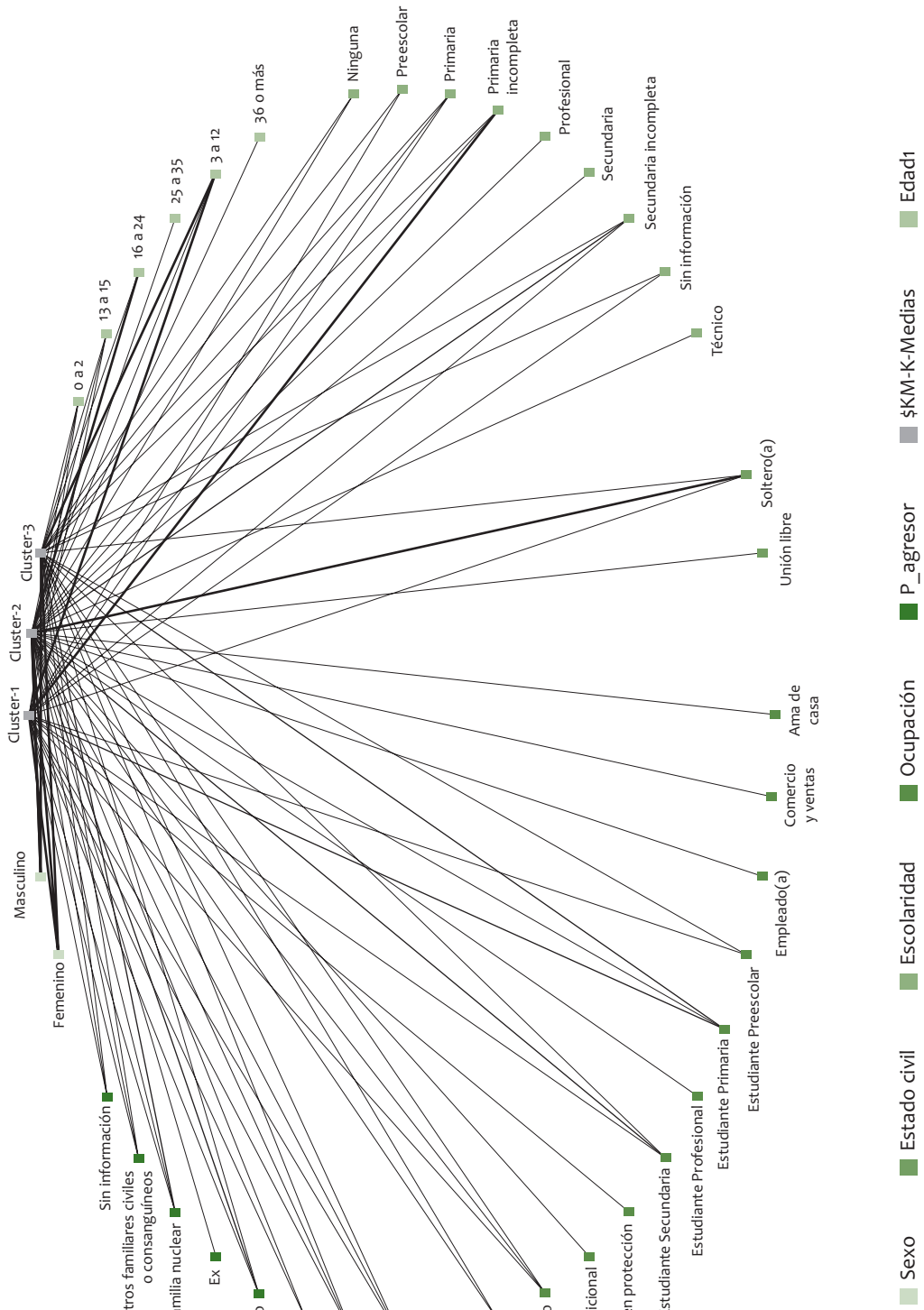
Regla 9 para CONOCIDO CON ALGÚN TRATO

si EDAD1 = 3 A 12 y ESCENARIO = HOGARES DE PROTECCIÓN SOCIAL y CIRCUNSTAN = OTROS entonces CONOCIDO CON ALGÚN TRATO

Regla 10 para CONOCIDO CON ALGÚN TRATO

si ULOCAL = ATENCIÓN AL MENOR y EDAD1 = 13 A 15 e IDENTIFICACIÓN = INDOCUMENTADO e INTERVALO1 = 06:01 a las 12:00 y ESCENARIO = VIVIENDA entonces CONOCIDO CON ALGÚN TRATO

Apéndice 9: Malla direccional MK-K-Medias - Medias sociodemográficas



Apéndice 10: Malla direccional MK-K-Medias - Medias contextuales