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

DISEÑO Y APLICACIÓN DE UNA METODOLOGÍA PARA ANÁLISIS DE NOTICIAS POLICIALES UTILIZANDO MINERÍA DE TEXTOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

DANIEL ALEJANDRO TORRES SILVA

PROFESOR GUÍA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
SEBASTIÁN RÍOS PÉREZ
JUAN VELÁSQUEZ SILVA

SANTIAGO DE CHILE
JUNIO 2013

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: DANIEL TORRES SILVA
FECHA: 27-06-2013
PROF. GUÍA: RICHARD WEBER H.

DISEÑO Y APLICACIÓN DE UNA METODOLOGÍA PARA ANÁLISIS DE NOTICIAS POLICIALES UTILIZANDO MINERÍA DE TEXTOS

En esta memoria se presenta el diseño y la aplicación de una metodología, basada en técnicas de minería de textos, para el procesamiento de grandes volúmenes de noticias que permita descubrir interesantes patrones en los datos para desarrollar un acabado análisis de la cobertura de distintas temáticas policiales y estudiar su relación con las estadísticas de casos reales de delincuencia.

Las noticias policiales han sido objeto de constante análisis, cuyo interés se debe a la probable distorsión que provocaría sobre los niveles de preocupación frente al delito en la población. Los artículos noticiosos poseen valiosa información que muchas veces no es explotada, dado que requiere de un proceso de análisis manual, intensivo en tiempo y recursos. Actualmente existen herramientas tecnológicas que permiten manejar crecientes volúmenes de datos, particularmente datos no estructurados como los textos, tomando cada vez más protagonismo la minería de textos en el descubrimiento de nuevo e interesante conocimiento.

La metodología de investigación propuesta se basa en las etapas del modelo de procesos CRISP-DM, para lo cual se debe integrar la comprensión de la naturaleza del problema, la comprensión y preparación de los datos, la construcción y evaluación de los modelos y los posteriores desarrollos a partir del conocimiento generado.

La aplicación de la metodología se realiza sobre un conjunto de noticias policiales en formato RSS recopiladas a partir de cuatro medios de prensa nacionales durante el segundo semestre del año 2011. Se logran identificar siete temáticas policiales dentro de las noticias, caracterizando cada una de ellas a partir de diferentes herramientas basadas en las palabras más relevantes. Se detecta que las distintas temáticas presentan diferentes niveles de cobertura entre sí, así como también según región y según medio de prensa. También se logra constatar una relativa proporcionalidad entre el número de noticias y el número de habitantes según región, mientras que la posible relación entre el número de casos reales y el número de noticias podría explicar una proporción importante de la variabilidad experimentada por los niveles de noticias para cada temática policial.

La metodología implementada permite cumplir exitosamente con los objetivos propuestos, facilitando la comprensión y manejo de las distintas interacciones entre las etapas involucradas en el proceso. Como trabajo futuro se plantea un sistema de monitoreo continuo de los niveles de cobertura policial en medios de prensa escritos, apoyándose en la propuesta de este trabajo.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

1. Introducción

1.1. Descripción del proyecto y justificación

La delincuencia ha sido señalada, a través de estudios de opinión[1] durante los últimos cinco gobiernos, como una de las principales problemáticas que la población espera mayor esfuerzo en su solución por parte de las autoridades. Debido a lo anterior, el tema de la preocupación por la delincuencia es continuamente monitoreado tanto por el gobierno (a través de la medición de la Encuesta Nacional Urbana de Seguridad Ciudadana, ENUSC) como por instituciones privadas (como el Centro de Estudios Públicos, CEP).

En la actualidad, la información sobre noticias abunda en los medios de comunicación de masas, existiendo cada vez más medios que se dedican exclusivamente al área tales como sitios web, canales de televisión, blogs, etc. Asimismo, nuevas tecnologías, como el formato RSS, surgen para satisfacer las crecientes necesidades específicas de los usuarios para acceder al contenido de forma más rápida y fácil.

En diversos estudios[2] con un enfoque sociológico, tanto nacionales como extranjeros, se ha analizado el efecto de los niveles de cobertura policial en los medios de prensa y su efecto dentro de una sociedad, principalmente como factor relevante en la distorsión presente en los niveles de percepción frente al delito. En varios de estos estudios se ha encontrado que los niveles de preocupación de las personas respecto al delito no estarían directamente relacionados con los niveles de casos delictuales reales ocurridos y/o con la experiencia personal como víctima, sino más bien podría deberse a factores adicionales como el tratamiento del tema policial en los medios de comunicación.

El manejo de grandes volúmenes de información puede transformarse en un gran desafío, por lo que es esencial contar con herramientas para enfrentar con rapidez y dinamismo esta tarea, especialmente cuando se trabaja con datos no estructurados como los documentos de textos. Un porcentaje importante de la información se encuentra en forma de texto, el que muchas veces no es aprovechado a cabalidad para su análisis debido a que su manejo implica cierto grado de complejidad adicional originado por la ausencia de estructura. Cada día surgen nuevas metodologías y herramientas que permiten un manejo eficiente de grandes volúmenes de datos no

[1] De acuerdo a los resultados obtenidos en la Encuesta CEP frente a la pregunta: ¿Cuáles son los tres problemas a los que debería dedicar el mayor esfuerzo en solucionar el gobierno?, disponible en: http://www.cepchile.cl/graficos_EncCEP/graf_evolProblemas.htm

[2] En la sección 2.1. de este informe se profundiza en dichos estudios.

estructurados, cuyo uso permite obtener nuevo y valioso conocimiento en diversas áreas de aplicación.

El análisis manual de noticias, que requiere el empleo de grandes cantidades de información en forma de texto, es un proceso lento, poco estructurado y propenso a errores, sumado a que dicho análisis debe realizarse diferido y acotado a la capacidad humana de respuesta, cuyos resultados dependen fuertemente de la experiencia del analista. Los datos contenidos en las noticias pueden transformarse en una valiosa fuente de información, por lo que resulta necesario automatizar las tareas involucradas en el procesamiento de datos con el fin de lograr un análisis más eficiente.

Las herramientas de minería de textos permiten colaborar en el análisis y visualización de grandes colecciones de datos no estructurados con gran desempeño, ayudando a incrementar la calidad de los resultados y reduciendo los tiempos para obtener interesante conocimiento a partir de datos en bruto. Una de las principales utilidades de estas herramientas es el descubrimiento de relaciones entre la información no evidente contenida en documentos de textos y una fuente externa de información.

El proyecto de investigación que se desarrolla en esta memoria pretende profundizar sobre el nivel de criminalidad presente en las noticias policiales. Para ello se trabaja en el diseño y la aplicación de una metodología que permita de forma eficiente y estandarizada realizar las distintas tareas que involucra el análisis de noticias. Se plantea el diseño de un mecanismo metodológico para el procesamiento de los datos contenidos en grandes colecciones de noticias (textos) y la extracción de útil conocimiento, con el objetivo de utilizar dicho conocimiento para un posterior análisis de los niveles de cobertura policial y su relación con variables de fuentes externas como las estadísticas de casos de delitos reales. La aplicación de variadas herramientas basadas en minería de textos será clave para identificar la presencia de distintas temáticas policiales tratadas en las noticias y la posterior caracterización de cada una de ellas.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general de este trabajo es diseñar y aplicar una metodología, basada en técnicas de minería de textos, que permita apoyar el procesamiento de grandes volúmenes de noticias para realizar un posterior análisis a partir de la cobertura de las distintas temáticas policiales detectadas en dichas noticias.

1.2.2. Objetivos específicos

- Estudiar el efecto y la importancia de las noticias policiales como factor relevante dentro de una sociedad a través de la revisión de diferentes estudios y teorías.

- Diseñar un mecanismo para la identificación de distintas temáticas policiales dentro de las noticias, aplicando posteriormente distintas herramientas para la caracterización de cada una de las temáticas.

- Analizar la distribución geográfica de las noticias policiales y estudiar las posibles relaciones entre el número de noticias y el número de casos reales según temática policial.

- Realizar un prototipo de una herramienta de apoyo para la visualización georreferenciada de la información más relevante obtenida.

1.3. Resultados esperados

Los principales resultados a obtener del diseño y aplicación de la metodología a realizar para el análisis de noticias policiales son:

- Identificación de diferentes temáticas dentro de las noticias policiales.

- Caracterización de las distintas temáticas policiales utilizando diferentes herramientas.

- Evaluación de la capacidad predictiva de un modelo para la clasificación de noticias dentro de un conjunto de temáticas policiales.

- Estudio de la distribución de las diferentes temáticas policiales dentro de las noticias para cada región.

- Estudio comparativo del número de noticias policiales y del número de casos de delitos reales por habitantes según región.

- Revisión de una posible relación entre el número de noticias policiales y el número de casos de delitos reales para cada temática policial identificada.

- Implementación de un prototipo de apoyo para la visualización de datos georreferenciados.

1.4. Alcances

El estudio pretende establecer una metodología para el análisis de noticias policiales basado en minería de textos cuya aplicación, en forma de experimentación, se realiza sobre un grupo de noticias y en un período de tiempo acotado.

Se utiliza como principal recurso a las noticias nacionales de carácter policial de cuatro medios de prensa chilenos distintos, que distribuyen su contenido en formato RSS. Dado que, en general, los distintos medios de prensa presentan diferentes líneas editoriales, la cobertura de cada una de las temáticas policiales por identificar puede tener variaciones considerables.

Las noticias son recopiladas durante seis meses, desde julio a diciembre del año 2011. Es posible que durante el período de estudio en el cual se realiza la recopilación de noticias pueda estar afectado por algún hecho noticioso de gran connotación, que pueda afectar la orientación de los resultados.

El análisis de las noticias policiales posee una fuerte orientación hacia el estudio del factor localización geográfica utilizando como recursos complementarios el número de casos reales de delitos y el número de habitantes por región, entendiendo que pueden existir otros factores relevantes.

2. Marco conceptual

2.1. Noticias policiales y su efecto social

Diversos estudios plantean que un alto nivel de cobertura y sensacionalismo presente en las noticias policiales afecta el sentido de seguridad y temor frente al crimen por parte de la población [21]. Es esperable que la percepción del delito esté correlacionada con las estadísticas delictivas, enlazando plano subjetivo y objetivo respectivamente. Cuando el temor al delito presenta un incremento notorio que no coincide con un cambio significativo en el número real de casos, los medios de comunicación son vistos muchas veces como uno de los probables causantes [13]. Diversos estudios se han realizado en torno a las noticias como factor relevante dentro de la sociedad, algunos de los cuales serán descritos brevemente en esta sección.

La investigación realizada en [21] consistió en un análisis cuantitativo y cualitativo de las noticias de delitos obtenidas a partir de cinco diarios nacionales durante un mes. Se distinguieron tres grandes grupos de noticias de delitos: el primer grupo formado por delitos de mayor connotación social, violencia intrafamiliar y drogas; el segundo grupo formado por noticias sobre instituciones y políticas públicas en seguridad ciudadana; y el tercer grupo referente a hechos de desorden social. Las noticias del primer grupo destacan por su distribución interna, en la que los delitos cometidos contra las personas tienen la extensa cobertura (casi el 50%) seguida por los delitos contra la propiedad, delitos relacionados a drogas y violencia intrafamiliar. Las principales conclusiones obtenidas en dicha investigación son las siguientes:

- Una de las principales características de las noticias policiales es el factor de dramatismo y sensacionalismo impreso en las historias, que muchas veces sirve como factor clave para captar al lector.

- Cierto tipo de noticias policiales tendría mayor cobertura basado en el impacto en los lectores que tendrían los aspectos dramáticos de un hecho. Este es el caso de las notas sobre homicidios que presentan una ocurrencia relativamente baja frente otros tipos de delitos, pero que es una de las temáticas más tratadas por los medios. Las noticias de delitos destacan en los medios debido a la alta importancia que le asignan, situando este tipo de noticias en los titulares, aunque esto no signifique una extensa cobertura del suceso.

- Las noticias policiales se centran en la cobertura del hecho dejando de lado muchas veces la evolución del mismo. Este tipo de noticias son entregadas en gran cantidad en un formato de información breve, generando una sensación de ocurrencia de gran número de delitos sin resolución, lo que no permite a los

lectores un acabado entendimiento del problema social de fondo. Además se observa cierta estigmatización de zonas específicas a la hora de informar hechos delictuales.

- Lo que informan los medios de información no necesariamente es lo que en realidad ocurre en una sociedad, dado que está sujeto muchas veces a criterios editoriales o de mercado, por lo que es natural encontrar que diferentes medios dan distinta cobertura a un mismo suceso.

En [13] se estudiaron noticias policiales durante tres meses, presentes en periódicos populares, para comparar la cobertura del delito entre la prensa chilena e inglesa, tomando una muestra de 140 notas para describir el caso chileno y una muestra de 173 notas para el caso inglés. Para ambos casos se detecta que los distintos tipos de delitos no son uniformemente reporteados, ya que más del 70% de las noticias se focalizaron en crímenes violentos, subestimando los delitos más comunes, lo que no corresponde a lo expresado por las cifras oficiales de casos reales. En el caso inglés la temática con mayor presencia en las noticias fue el homicidio, mientras que para el caso chileno la mayor cobertura fue para el robo con violencia.

En [74] se estudian las noticias de un diario español, las encuestas de opinión y los datos oficiales sobre delincuencia, basado en el supuesto de que la imagen sobre la criminalidad que forma una persona se fundamenta principalmente en su experiencia directa como víctima, la experiencia como víctima de sus cercanos y las noticias difundidas por los medios de prensa. Se determina que en las noticias policiales predominan los sucesos de homicidios y asesinatos, dejando un menor porcentaje a los casos de robos, lo cual no corresponde a la proporción entre los casos reales de estos tipos de delitos. Se señala además que los índices más altos de preocupación al delito se caracterizan por los relatos de sucesos más cruentos y no necesariamente destacan por el número de noticias de delitos en el mismo período. Dentro de las conclusiones se generales se encuentran:

- Los medios de comunicación pueden alterar la visión de la criminalidad real, incrementando o disminuyendo la cobertura de ciertas temáticas, lo cual puede afectar seriamente la efectividad de las políticas de prevención del delito.

- La influencia de los medios de comunicación corresponde a una explicación bastante razonable, aunque no exclusiva, de los incrementos de los niveles de preocupación del delito, considerando que el número de casos reales de delitos tiende a tener un comportamiento bastante estable en el tiempo.

En [31] se analizan cuatro noticieros de distintos canales chilenos de televisión y se confirma que el tema con mayor cobertura en la agenda noticiosa es la seguridad ciudadana, durante diez de los doce meses analizados.

En [25] se estudiaron 105 ediciones de informativos televisivos españoles sobre noticias de delitos en un lapso de casi tres meses, las que fueron transcritas completamente. A partir del análisis se establece una lógica mediática que resalta las noticias negativas, siendo recurrentes en su estructura los hechos de robos, asesinatos y violaciones. En general, se encuentra que en los espacios informativos destacan los detalles, muchas veces morbosos, de los sucesos concretos por sobre los elementos que facilitarían su comprensión, probablemente para aumentar el impacto emocional de los hechos. Dado lo anterior, se privilegiaría los objetivos de entretener y emocionar por sobre informar.

En [5] se constató que la recordación espontánea del último hecho delictual está fuertemente mediatizada y se formula la hipótesis de que los medios no son tan influyentes en el temor de una persona a ser víctima de un delito, sino más bien afectaría en la percepción de la delincuencia como problema social a nivel global, lo que concuerda con estudios internacionales más recientes [78].

En [37] se concluye que la cobertura mediática de los crímenes violentos tiene efectos en la sensación de temor solo para ciertos crímenes específicos. Por otro lado, considera que el tipo de medio que informa la noticia delictual, ya sea escrito o televisivo, tendría un impacto similar sobre la sensación de temor.

Dos de las principales teorías sobre la influencia de los medios de comunicación en la percepción del crimen en las personas son: la teoría de la agenda-setting y la teoría del cultivo. La teoría de la agenda-setting [54] [55] analiza la influencia mediática en la inclusión de la criminalidad en la agenda del público, concluyendo que el protagonismo en la agenda pública se debe gran parte a la agenda mediática y no a los cambios en las tasas de criminalidad. En la teoría del cultivo [32] [33] se estudia la influencia de los medios de comunicación en la percepción del público sobre hechos de violencia en la ciudad y en el temor a ser objeto de violencia, concluyendo que el espacio mediático contenía muchas más violencia de la existente en hechos concretos y que tal desproporción hace creer que existe mayor violencia de la real y al aumento del temor de ser víctima de un hecho violento. De esta forma una entrega repetitiva de contenido específico puede “cultivar” en las personas imágenes distorsionadas de la realidad de su entorno.

2.2. Proceso KDD

El proceso KDD (Knowledge Discovery in Databases) [27] está definido como el “proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensibles en los datos”. El término proceso implica que KDD

involucra muchos pasos y el término no trivial implica que no se tratan de cálculos sencillos.

El proceso KDD es interactivo e iterativo, involucra numerosos pasos con muchas decisiones que deben ser tomadas por el usuario [27] [28]:

- Selección: consiste en la creación de una conjunto de datos objetivo, seleccionando un conjunto de datos o focalizándose en un subconjunto de variables o un muestreo de los datos, sobre el cual el descubrimiento de conocimiento será realizado. En este paso las distintas fuentes de datos pueden ser combinadas.

- Pre-procesamiento: tiene por objetivo la limpieza de datos con el fin de obtener datos consistentes. Las operaciones básicas incluyen la remoción de ruido si es necesario y estrategias para el manejo de los campos de datos vacíos.

- Transformación: consiste en la transformación de los datos usando reducción de dimensionalidad o métodos de transformación. Los datos son transformados en formas apropiadas para la minería de datos y/o se seleccionan los atributos más útiles capaces de representar los datos dependiendo de las metas propuestas.

- Data Mining: es un proceso donde los métodos de inteligencia son aplicados con el fin de realizar una búsqueda y extracción de patrones de datos de interés en una forma particular de representación, por ejemplo clasificación o clustering.

- Interpretación/Evaluación: consiste en la interpretación y evaluación de los patrones encontrados. Se identifican los patrones realmente interesantes basados en alguna medida de interés.

El proceso KDD debe ser precedido por el desarrollo de un entendimiento del dominio de aplicación (área específica donde se aplicará el proyecto de minería de datos) y la correcta identificación de las metas desde el punto de vista del usuario final [7] [27]. Una vez finalizado el proceso KDD, el conocimiento descubierto puede ser usado directamente, incorporado a otros sistemas para futuras acciones o simplemente documentado y reportado a las partes interesadas [28].

KDD puede involucrar importantes iteraciones y contener bucles entre pares de etapas [27]. La mayor parte del trabajo en el proceso KDD está enfocado en la etapa del data mining, aunque el resto de las etapas son igualmente importantes para la exitosa aplicación de KDD en la práctica.

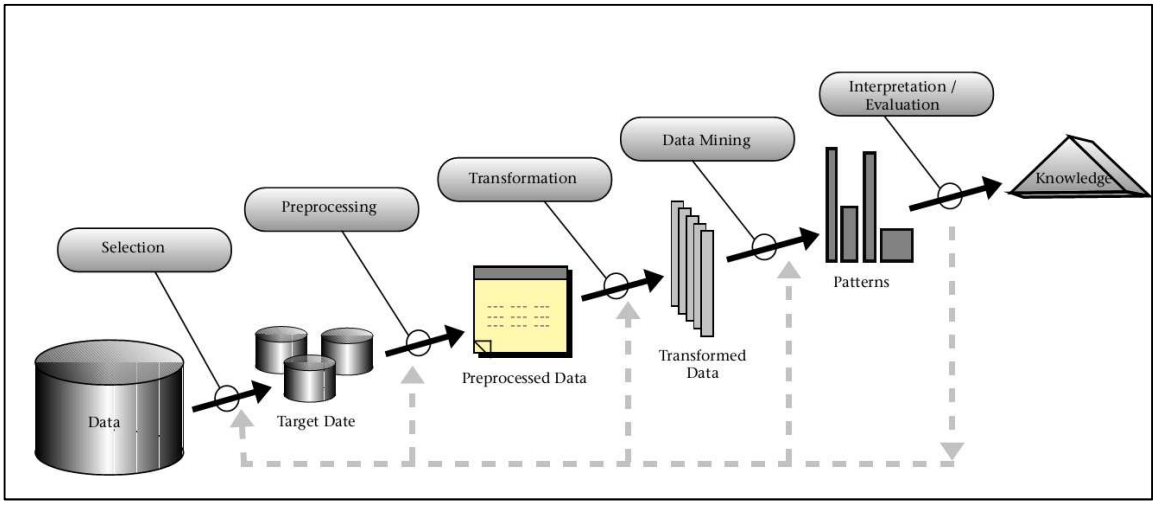

Figura 1: Proceso KDD
Fuente: [27]

2.2.1. CRISP-DM

El modelo CRISP-DM (Cross Industry Standard Process for Data Mining) [16] es una metodología que proporciona una descripción del ciclo de vida del proyecto de minería de datos. El modelo consiste en un ciclo de seis fases:

- Comprensión del ámbito de aplicación: esta fase se enfoca en la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva no técnica, para luego convertir este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

- Comprensión de los datos: esta fase comienza con la recopilación inicial de datos y continúa con las actividades que permiten familiarizarse con los datos y verificar su calidad.

- Preparación de datos: esta fase cubre todas las actividades necesarias para construir el conjunto de datos que será utilizado en las herramientas de modelado a partir de los datos en brutos iniciales. Las tareas incluyen selección de atributos junto con limpieza, integración y formateo de datos.

- Modelado: en esta fase varias técnicas de modelado son seleccionadas y aplicadas. También se realiza el diseño para la posterior evaluación del modelo.

- Evaluación: se evalúan los resultados finales obtenidos del modelo y se revisan los pasos del proceso realizado. Luego se compara el modelo obtenido con los objetivos inicialmente planteados.

- Desarrollo: en esta fase el conocimiento obtenido es organizado y presentado en el modo en el que el usuario final pueda usarlo. Se establece una planificación de la monitorización y del mantenimiento del proceso, se genera los informes finales y se realiza la revisión del proyecto.

En la Figura 2 se muestra el proceso CRISP-DM, en donde la secuencia de las fases no es rígida, ya que el movimiento entre fases diferentes es siempre requerido. Las flechas indican las dependencias más importantes y frecuentes entre fases.

CRISP-DM puede ser visto como una implementación del proceso KDD, guiando a los usuarios en la puesta en práctica de data mining en sistemas reales [7]. En el Cuadro 1 se puede observar las correspondencias entre las etapas entre KDD y CRISP-DM, donde la etapa de comprensión del ámbito de aplicación puede ser identificada con el desarrollo del entendimiento del dominio de aplicación, el conocimiento previo relevante y las metas de los usuarios finales, mientras que la etapa de desarrollo puede ser identificada con la consolidación a través de la incorporación del conocimiento al sistema.

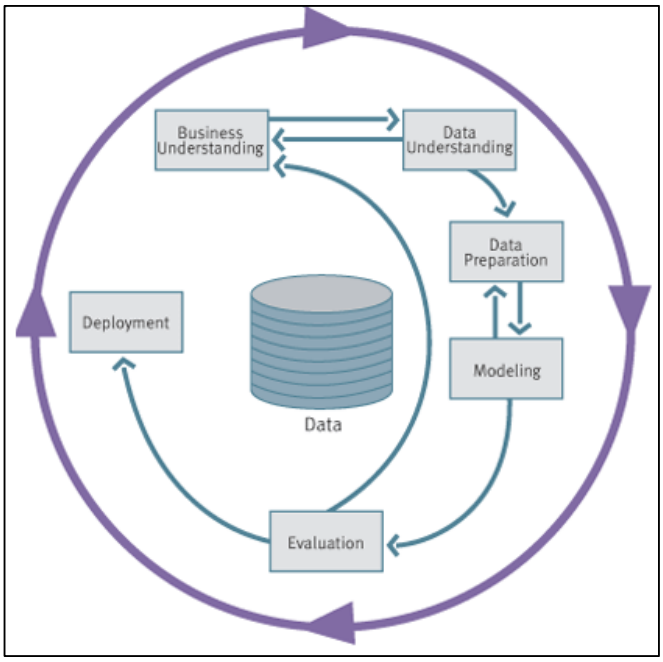

Figura 2: Etapas del proceso CRISP-DM
Fuente: [16]

KDD	CRISP-DM
Pre KDD	Comprensión del Ámbito de Aplicación
Selección	Comprensión de los Datos
Pre Procesamiento	
Transformación	Preparación de los Datos
Data Mining	Modelado
Interpretación/Evaluación	Evaluación
Post KDD	Desarrollo

Cuadro 1: Resumen correspondencias entre KDD y CRISP-DM. Fuente: [7]

2.3. Text mining

Text mining puede ser definido [43] como “la aplicación de algoritmos y métodos de los campos del aprendizaje de máquina y la estadística sobre los textos con el objetivo de encontrar patrones útiles”. Text mining es una herramienta empírica que tiene la capacidad de identificar nueva información o patrones significativos que no son evidentes a partir de una colección de documentos [35] [82].

En principio se puede utilizar cualquiera de los métodos de clasificación estándar usado en data mining para la aplicación en text mining, pero el conjunto de datos proveniente de los documentos de textos necesita un tratamiento previo [12]. Para aplicar text mining a grandes colecciones de documentos es necesario realizar un pre-procesamiento de los documentos de textos y almacenar la información de una forma estructurada [43].

El primer paso en text mining es recopilar los documentos relevantes. Luego la principal tarea es limpiar las muestras y asegurar la calidad, debido a que se pueden encontrar diversos formatos entre los documentos. La transformación del texto a datos numéricos permite que los datos cambien a una codificación clásica de data mining. Por lo que la presentación de los datos para data mining y text mining son solo diferentes en una presentación inicial [84]. Existen muchas variantes de representación de documentos, la mayoría de los enfoques están basados en que las palabras son atributos y los documentos son muestras, formando el conjunto de datos que permitirá desarrollar los distintos métodos de aprendizaje.

Las fuentes de conocimiento externas pueden beneficiar enormemente varios de los elementos que componen un sistema de text mining, entregando limitaciones o información adicional sobre los conceptos que se encuentran a través de la aplicación de métodos de aprendizajes, así como servir en la etapa de validación de los resultados [29].

2.3.1. Representación de documentos

La representación de un documento de texto se basa en las palabras, modelo conocido como Vector Space Model o “bag of words”. Con esta representación un documento es considerado simplemente como una colección de palabras que ocurren al menos una vez. El orden de las palaras, la combinación en las cuales ellas ocurren, la estructura gramatical, la puntuación y el significado de las palabras son todos ignorados [12].

Al conjunto de atributos típicamente se le llama diccionario [84]. Existen dos enfoques para desarrollar un diccionario: local y global. El enfoque de diccionario local se puede utilizar cuando se conoce a priori las distintas categorías a las que pueden pertenecer los documentos, elaborando para cada categoría un diccionario distinto, solo con las palabras que aparecen en los documentos clasificados en la categoría específica. En el enfoque de diccionario global se incluyen todas las palabras que ocurren al menos una vez en cualquiera de los documentos. El primer enfoque logra que cada diccionario sea relativamente pequeño pero tiene un alto costo en tiempo para su construcción.

En general, los documentos de texto tienen un gran número de atributos y la mayoría de ellos ocurren con una baja frecuencia. En muchas circunstancias, se quiere trabajar con un diccionario más pequeño y la información sobre la frecuencia de ocurrencia de las palabras puede muy útil para la reducción del diccionario y puede mejorar el desempeño predictivo para algunos métodos. Tanto las palabras más frecuentes como las menos frecuentes podrían ser descartadas.

Una vez determinado que el número final de atributos es N, se puede representar los términos en el diccionario en algún orden arbitrario como $t_1,t_2,\ldots,t_N$. Así se puede representar el i-ésimo documento como un conjunto ordenado de N valores: (X_{i1}, X_{i2}, ... , X_{iN}). El conjunto de vectores que representan todos los documentos en consideración es llamado un Vector Space Model. En general, se puede decir que el valor X_{ij} es una medida ponderada de la importancia del j-ésimo término t_j en el documento i [12].

Cada elemento del vector representa un término (una palabra o un grupo de palabras) de la colección de documentos, es decir, el tamaño del vector está definido por el número de palabras de la colección completa de documentos.

Vector Space Model analiza eficientemente grandes colecciones de documentos, a pesar de su estructura simple y sin el uso de información semántica. Este modelo fue introducido en [71] principalmente para information retrieval, pero actualmente es aplicado extensamente en text mining [43].

Existen varias formas de representar la importancia ponderada de un término dentro de un documento: frecuencia, binaria y tf-idf [12] [43].Una de las formas más comunes

de calcular la importancia ponderada es contar el número de ocurrencia para cada término en un documento dado. Una representación binaria consiste en que el valor 1 indica que el término está presente en el documento, de lo contrario, el valor 0 indica la ausencia del término. Una forma más compleja de calcular la importancia ponderada es llamada tf-idf (term frequency inverse document frequency), la cual combina la frecuencia de un término con una medida de rareza del término en el conjunto completo de documentos.

2.3.2. Tokenización

El primer paso para el manejo de texto es separar el stream de caracteres para llevarlo a palabras o, más precisamente, a tokens. El proceso de tokenización es altamente dependiente del lenguaje sobre el cual se está trabajando y las reglas específicas de este, por lo que esta tarea resulta trivial para una persona familiarizada con la estructura del lenguaje, mientras que para un programa computacional puede convertirse en todo un reto [84].

Para identificar las distintas palabras (o tokens) es necesario primero definir los delimitadores de los tokens, los que generalmente corresponden a los signos de puntuación y otros caracteres distintos a las letras del alfabeto [43]. Luego los delimitadores de tokens se separan de las palabras y son reemplazados por un espacio blanco simple [10]. De esta forma cada palabra queda separada por un espacio blanco simple y facilita la tokenización.

2.3.3. Stopwords

Stop words es un método de filtración que consiste en remover palabras comunes que resultan inútiles para la caracterización de un documento [12].

La idea de la aplicación de un lista de stop words es remover palabras que ocurren muy poco o no contienen información útil, como artículos, pronombres, conjunciones, preposiciones, etc. Las palabras que aparecen en la mayoría de los documentos aportan poca información para distinguir los distintos tipos de documentos, así como las palabras que ocurren muy raramente es probable que no tengan una relevancia estadística y pueden ser removidas del diccionario [43].

Al remover las palabras de la lista stop words se puede lograr una reducción significativa del tamaño del diccionario [84], pero no existe una lista fija de stop words que sea universalmente utilizada [12].

2.3.4. Stemming

Luego de que el documento de texto ha sido separado en una secuencia de tokens, el siguiente posible paso es convertir cada tokens a una forma estandarizada, proceso llamado stemming [84].

La aplicación de stemming se basa en la observación de que las palabras en los documentos a menudo tienen muchas variantes morfológicas. Las palabras que tiene una misma raíz lingüística pueden ser tratadas como una única palabra, la cual entrega probablemente una mejor descripción del contenido del documento, lo que podría no ocurrir si se utilizara cada palabra individualmente [12].

El objetivo de stemming es reconocer los conjuntos de palabras que pueden ser tratadas como equivalentes. Muchas veces no hay necesidad para mantener el singular y plural de una misma palabra, así como los verbos pueden ser almacenados es su forma infinitiva. También se puede extender el concepto hacia los sinónimos.

Algunos de los efectos de la aplicación de stemming es la reducción del número total de atributos dentro del texto (o reducción del tamaño del diccionario [43]) y el incremento de la frecuencia de ocurrencia de algunos atributos [84].

Tal como para stop words, no hay un algoritmo de stemming que sea universalmente usado, donde el idioma tiene un papel clave. Por otro lado la aplicación de stemming debe ser implementado con cautela para no eliminar palabras del diccionario que puedan resultar relevantes, dado que no se considera la semántica de las palabras [12].

2.3.5. Similitud entre documentos

Una medida natural para comparar dos documentos es la similitud. Dos documentos son similares cuando comparten las mismas palabras, así que mientras más palabras se compartan más similares serán los documentos. Algo muy relevante de la medida de similitud es que las palabras que ocurren en un documento pero no en los otros son ignoradas (lo que no ocurre en las medidas de distancia convencionales), lo que ayuda a resolver el problema de la representación de un documento como un vector disperso [84].

La distancia euclidiana es usada en problemas de clustering y de clasificación. La medida de distancia euclidiana entre dos documentos de textos, d_a y d_b, representados por sus vectores de términos $\overrightarrow{\boldsymbol{t_a}}$ y $\overrightarrow{\boldsymbol{t_b}}$ respectivamente, está definida como:

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = \left(\sum_{t=1}^{m} | w_{t,a} - w_{t,b} |^2 \right)^{1/2}$$

donde el conjunto de términos es T={t_1,…t_m} y $w_{t,d}$ es una ponderación del término t en el documento d [44]. Distancias métricas como la euclidiana no son las más apropiadas para dominios altamente dimensionales y dispersos [76], como lo son los documentos de texto.

La medida de similitud de coseno entre dos documentos de textos, d_a y d_b, representados por sus vectores de términos $\overrightarrow{t_a}$ y $\overrightarrow{t_b}$ respectivamente, está definida como [39] [75]:

$$SIM_c(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{| \overrightarrow{t_a} | \ | \overrightarrow{t_b} |}$$

donde $\overrightarrow{t_a} \cdot \overrightarrow{t_b}$ es el producto punto vectorial, definido como $\sum_{t=1}^{m}(w_{t,a}\, w_{t,b})$, y la norma $| \overrightarrow{t_a} |$ en el denominador está definida por $| \overrightarrow{t_a} | = \sqrt{\overrightarrow{t_a} \ \overrightarrow{t_a}}$, con el conjunto de términos T={t_1,…t_m} y $w_{t,d}$ es una ponderación del término t en el documento d. Similitud de coseno es una medida que captura exitosamente la similitud implícita entre documentos y es la medida de similitud más utilizada para tareas de procesamiento de textos [29] [76] [84].

La similitud de coseno es una medida de similitud y no de distancia, pero es posible aplicar una simple transformación para convertir la medida de similitud en un valor de distancia [44]. Dado que la similitud de coseno está limitada por [0,1] y es monótona, se toma D=1-SIM como el correspondiente valor de distancia.

Para datos con alta dimensionalidad como es el caso de documentos de textos, la similitud de coseno ha mostrado ser una medida significativamente superior a la distancia euclidiana [44] [76]. La implicancia de esto es que la dirección de un vector de documentos es más importante que su magnitud [89].

2.3.6. Selección de atributos

La representación de un texto como una bolsa de palabras (Vector Space Model) plantea dos problemas serios [51] [77] [86]: alta dimensionalidad del espacio de atributos y la inherente dispersión de los datos, dado que los atributos pueden ser miles para una colección moderada de documentos de textos. Debido a lo anterior, los algoritmos de aprendizaje que serán aplicados sobre los datos pueden presentar un muy bajo desempeño [50].

Dada la negativa influencia de la alta dimensionalidad y la dispersión de los datos, resulta fundamental aplicar una reducción de la dimensionalidad del espacio de atributos [36]. Existen dos técnicas utilizadas usualmente para tratar la reducción de dimensionalidad: extracción de atributos y selección de atributos.

La extracción de atributos es un proceso que extrae un conjunto de nuevos atributos desde los atributos originales [51]. Los nuevos atributos generados son producto de alguna combinación lineal o no lineal de los atributos originales [77]. La principal desventaja que puede presentar la aplicación de métodos de extracción de atributos es que al generar nuevos atributos, estos no tengan un significado claro [20] [50], debido a lo cual los resultados derivados de métodos de aprendizaje sobre los datos serán complejos de interpretar.

La selección de atributos es un proceso que elige un subconjunto del conjunto de atributos originales de acuerdo a algún criterio. Los atributos seleccionados conservan su significado original y entregan un mejor entendimiento de los datos y los procesos de aprendizaje. [50] [51] [77] La selección de atributos además mejora el desempeño de la capacidad predictiva de los atributos conservados y ayuda a conseguir ahorros en los requerimientos de almacenaje de datos y en los tiempos de ejecución de los procesos [24].

Dentro de los algoritmos de selección de atributos destacan aquellos que ordenan todos los atributos basados en alguna medida numérica y seleccionan un subconjunto de atributos utilizando dicha medida. Este tipo de algoritmos son ampliamente utilizados debido a su simplicidad y éxito empírico [36] [77].

2.3.6.1. Selección supervisada de atributos

Cuando el método de selección de atributos requiere la información de la clase a la cual pertenece un documento se trata de un método supervisado, en caso contrario se trata de un método no supervisado [51].

En los métodos de selección supervisada de atributos se debe determinar los atributos más correlacionados con las distintas clases [20]. Algunos métodos de selección supervisada de atributos han sido exitosamente utilizados para la clasificación de textos [50] [86]. Chi-Cuadrado (CHI) e Information Gain (IG) son formas distintas de medir la correlación entre los términos y las categorías (o clases) [2].

El estadístico Chi-Cuadrado mide la falta de independencia entre un término t y una clase c [2] [86], tomando el valor cero si t y c son independientes. Sea A el número de veces que el término t y la clase c co-ocurren, B es el número de veces que el término t ocurre sin que ocurra la clase c, C es el número de veces que la clase c ocurre sin que

ocurra el término t, D es el número de veces que no ocurre la clase c ni el término t y N es el número total de documentos. El estadístico Chi-Cuadrado queda definido por:

$$\boldsymbol{CHI}(\boldsymbol{t},\boldsymbol{c}) = \frac{\boldsymbol{N} \times (\boldsymbol{AD} - \boldsymbol{CB})^{\boldsymbol{2}}}{(\boldsymbol{A}+\boldsymbol{C}) \times (\boldsymbol{B}+\boldsymbol{D}) \times (\boldsymbol{A}+\boldsymbol{B}) \times (\boldsymbol{C}+\boldsymbol{D})}$$

Luego se puede calcular el valor de Chi-Cuadrado asociado a cada término [51], combinando los valores de las distintas categorías específicas:

$$\boldsymbol{CHI}(\boldsymbol{t}) = \frac{\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}} \boldsymbol{CHI}(\boldsymbol{t},\boldsymbol{c_i})}{m}$$

donde m es el número total de clases en el conjunto de documentos.

Information Gain mide el número de bits de información obtenida para predicción de categorías a través del conocimiento de la presencia o ausencia de un término en un documento [86]. Mientras mayor sea el valor de information gain del término t, mejor será el poder discriminatorio del término t [2]. Sea $\{\boldsymbol{c_i}\}_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}$ el conjunto de categorías, la ecuación de Information Gain está dada por:

$$\boldsymbol{IG}(\boldsymbol{t}) = -\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}} \boldsymbol{p}(\boldsymbol{c_i}) \log \boldsymbol{p}(\boldsymbol{c_i}) + \boldsymbol{p}(\boldsymbol{t}) \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}} \boldsymbol{p}(\boldsymbol{c_i}|\boldsymbol{t}) \; \boldsymbol{log} \; \boldsymbol{p}(\boldsymbol{c_i}|\boldsymbol{t}) + \boldsymbol{p}(\bar{\boldsymbol{t}}) \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}} \boldsymbol{p}(\boldsymbol{c_i}|\bar{\boldsymbol{t}}) \; \boldsymbol{log} \; \boldsymbol{p}(\boldsymbol{c_i}|\bar{\boldsymbol{t}})$$

Una vez calculado el valor de Chi-Cuadrado e Information Gain de cada término, se remueven aquellos atributos que tengan un valor menor a un umbral mínimo predeterminado.

2.3.6.2. Selección no supervisada de atributos

En la mayoría de los casos reales la información sobre la clase de cada documento es desconocida, por lo que métodos de selección no supervisada de atributos debe ser aplicada. Al aplicar métodos de selección no supervisada de atributos se debe seleccionar una mayor cantidad de atributos que en el caso supervisado, debido a que, en general, se obtiene una eficiencia más baja para el caso no supervisado. Debido a lo anterior, los métodos de selección no supervisada se deben aplicar de forma más cuidadosa [51].

Algunos métodos no supervisados de selección de atributos han sido exitosamente utilizados para la clustering de textos. A partir de experimentos realizados [50] se obtiene, para cuatro métodos de selección de atributos no supervisados (DF, TC, TVQ y TV), que la selección de atributos puede mejorar la eficiencia y exactitud del clustering de documentos y que al menos un 70% de los atributos pueden ser removidos sin pérdida de desempeño. Por otro lado hay que tener en cuenta que una deficiente

selección de atributos puede reducir en un grado importante el desempeño del algoritmo de aprendizaje que se utilizará [24].

Document Frequency (DF) es el número de documentos en los cuales un término ocurre al menos una vez en el conjunto de documentos. Se calcula el valor de DF para cada atributo y se eliminan aquellos que tengan un DF menor a un valor predeterminado. El supuesto básico es que los términos que ocurren en muy pocos documentos no resultan relevantes para la predicción de una categoría [50]. Este es el criterio más simple para la selección de atributos y es fácilmente aplicable a grandes conjuntos de datos con complejidad computacional lineal [51]. A pesar de su simplicidad, ha demostrado su efectividad a la par con técnicas más avanzadas de categorización de textos [73] [77]. En experimentos realizados en [86] se determinó que mediante el uso de DF es posible reducir la dimensionalidad del espacio de atributos en un factor de 10 sin pérdida de efectividad.

DF es utilizado ampliamente antes de aplicar tanto algoritmos de aprendizaje supervisados como no supervisados. Se encuentra una fuerte correlación entre los valores DF, IG y CHI para un atributo [30], siendo IG y CHI los métodos más efectivos en los experimentos realizados. Esto indica que el DF, el método más simple con el menor costo computacional, puede ser tan confiable como IG o CHI cuando los costos computacionales para obtener esas medidas son mucho más altos.

Term Contributtion (TC) [51] ordena los atributos por su contribución a la similitud entre los documentos en el conjunto de datos. TC se calcula con la siguiente ecuación:

$$\boldsymbol{TC}(\boldsymbol{t_k}) = \sum_{\boldsymbol{i,j\ ,i\neq j}} \mathrm{TF\ IDF}(\boldsymbol{t_k}, \boldsymbol{D_i}) * \mathrm{TF\ IDF}\left(\boldsymbol{t_k}, \boldsymbol{D_j}\right)$$

donde TF-IDF (Term frequency – Inverse document frequency) está definido por $\mathrm{TF\ IDF}\left(\boldsymbol{t_k}, \boldsymbol{D_j}\right) = \boldsymbol{TF_{kj}} * \mathbf{log}(\frac{\boldsymbol{N}}{\boldsymbol{DF_j}})$, con N el número de documentos, $\mathrm{TF_{kj}}$ el número de veces que el término $\boldsymbol{t_k}$ aparece en el documento j dividido por el número total de palabras contenidas en dicho documento y $\mathrm{DF_k}$ es el Document Frequency del término $\mathrm{t_k}$.

El método de Term Variance Quality (TVQ) es introducido en [24] y está basado en la varianza de la frecuencia de cada término. El método TVQ ha demostrado éxito reduciendo a solo el 15% de la dimensión original [48]. El valor de TVQ de un término esta medido como:

$$\boldsymbol{TVQ}(\boldsymbol{t_i}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{N}} \boldsymbol{f_{ij}}^{\mathbf{2}} - \frac{\mathbf{1}}{\boldsymbol{N}}\left(\sum_{\boldsymbol{j=1}}^{\boldsymbol{N}} \boldsymbol{f_{ij}}\right)^{\mathbf{2}}$$

donde N es el número de documentos en los cuales el término $\boldsymbol{t_i}$ ocurre al menos una vez y $\boldsymbol{f_{ij}}$ es la frecuencia del término $\boldsymbol{t_i}$ en el documento j, con $\boldsymbol{f_{ij} \geq 1,\ j = 1, \ldots, N.}$

Term Variance (TV) se calcula como la varianza de un término en todo el conjunto de documentos. De esta forma un término que aparece en muy pocos documentos o tiene una distribución uniforme entre los documentos tendrá entonces un bajo valor de TV. Term Variance se calcula como [50]:

$$\boldsymbol{TV(t_i) = \sum_{j=1}^{N} \left(f_{ij} - \bar{f}_i\right)^2}$$

donde $\boldsymbol{f_{ij}}$ es la frecuencia del término $\boldsymbol{t_i}$ en el documento j y $\boldsymbol{\bar{f}_i}$ es la media aritmética de la frecuencia del término $\boldsymbol{t_i}$ entre todos los documentos.

A partir de experimentos realizados en [50] [51] [77] se obtiene la siguiente relación del desempeño de los distintos métodos de selección de atributos no supervisados: TC > TV > TVQ > DF. TC es el método de selección de atributos no supervisados preferido para clustering de textos.

2.3.7. Clasificación de documentos

Clasificación es el proceso de encontrar un modelo (o función) que describa y distinga clases de datos [39]. Aplicado a text mining, el objetivo es encontrar la correcta clase para cada nuevo documento, dada una colección de documentos de textos y un conjunto de categorías [29].

El problema de clasificación está definido como sigue [2] [84]: se tiene un conjunto de documentos de entrenamiento D={d_1,...,d_N}, cada documento esta etiquetado con una clase de un conjunto de k valores diferentes {1,...,k}. Con el conjunto de entrenamiento se construye el modelo de clasificación, el cual relaciona los atributos de cada documento con una clase. En términos matemáticos, la clasificación de documentos corresponde a una función que asigna documentos a una clase, f: w→c donde w es un vector de atributos y c es una clase particular. En términos estadísticos el proceso consiste en la extracción de una muestra desde una población, se aprende a partir de ella, y luego se aplica el modelo construido a nuevos ejemplos no etiquetados extraídos de la misma población.

La clasificación de documentos es un proceso de aprendizaje supervisado porque al proceso se le indica previamente a que clase pertenece cada documento, la cual puede provenir de algún mecanismo externo como un feedback humano [82]. Una vez que los documentos son representados como vectores numéricos y completado la etapa de pre-procesamiento, se pueden utilizar directamente para text mining la mayoría de los modelos de clasificación utilizados en data mining estándar [2]. Entre los

modelos más ampliamente utilizados, que se ajustan a las características de los textos, destacan K-nearest neighbors, métodos logísticos y métodos probabilísticos [84].

La clasificación de textos puede ser vista como un proceso de dos pasos [39]. En el primer paso se llama de aprendizaje (o entrenamiento) en el cual se construye el clasificador a partir del análisis de un subconjunto de documentos cuya clasificación asociada es conocida previamente. En el segundo paso, se evalúa la exactitud de predicción del clasificador construido, para lo cual se le aplica el modelo encontrado a un subconjunto de documentos de testeo (independiente del conjunto de entrenamiento) y se le asigna una clase a cada documento, para finalmente comparar esta asignación con la clase original de los documentos de testeo.

La clasificación de documentos puede ser single-label o multi-label [29]. En el primer caso cada documento pertenece exactamente a una clase y en el segundo caso cada documento puede estar asociado a más de una clase.

El éxito en la clasificación depende en gran parte de si es posible hallar patrones de palabras en distintos subconjuntos de documentos, dichos patrones se forman cuando se producen combinaciones de palabras en una clase y no en el resto.

2.3.7.1. Clasificador Naive Bayes

El clasificador Naive Bayes es un clasificador probabilístico basado en el supuesto de que el valor de un atributo en una clase dada es independiente de los valores de los otros atributos, lo que permite hacer simplificaciones importantes en los cálculos involucrados [39]. Este clasificador es uno de los métodos más simples [2] y fáciles de implementar [82], además ha mostrado un buen desempeño en la práctica y es uno de los más utilizados.

Para clasificación de documentos, Naive Bayes modela la distribución de los documentos en cada clase usando un modelo probabilístico bajo la hipótesis de independencia de distribución de los distintos términos [2]. La idea básica de este clasificador es usar las probabilidades condicionadas de los términos y de las clases para estimar las probabilidades de que un documento pertenezca a una categoría [82].

El algoritmo de Naive Bayes calcula la probabilidad de que un documento d_i pertenezca a la clase c_k a través del teorema de Bayes [29]:

$$\boldsymbol{P(c_k|d_i) = \frac{P(d_i|c_k)\,P(c_k)}{P(d_i)}}$$

donde la probabilidad $P(d_i)$ no necesita ser calculada pues esta es constante para todas las categorias y para calular la probabilidad $P(d_i|c_k)$ primero se debe utilizar la

representación de un documento como un vector de atributos $d_i=(w_{i1},\ldots,w_{in})$ y hacer uso del supuesto de independencia de la distribución de los atributos dado una clase:

$$\boldsymbol{P(d_i|c_k)} = \prod_{j=1}^{n} \boldsymbol{P(w_{ij}|c_k)}$$

De esta forma se calcula la probabilidad de que un documento pertenezca a cada clase y se elige la clasificación que tenga el valor más alto.

Existen dos tipos de modelos de Naive Bayes utilizados para clasificación de documentos [12]. El primero de ellos es el modelo multivariado Bernoulli, en el cual se utiliza la presencia o ausencia de las palabras en la representación de un documento. El segundo de ellos es el modelo multinomial, en el cual se utiliza la frecuencia de las palabras en la representación de un documento.

2.3.7.2. Clasificador k-nn

K-nearest neighbours (k-nn) es un clasificador basado en la proximidad entre los objetos, cuya idea principal es que los documentos de una misma clase están bastantes cercanos con respecto a alguna medida de similitud [2].

Métodos como k-nn son llamados lazy learner, dado que no se construye ningún tipo de abstracción (modelo) a partir de los datos de entrenamiento, posponiendo todo el trabajo hasta cuando un nuevo objeto debe clasificarse [29], calculando la similitud entre el nuevo documento y los documentos de entrenamiento. En la etapa de entrenamiento solo se almacenan las representaciones de los documentos con sus respectivas clases.

El algoritmo de clasificación de k-nn consiste en los siguientes pasos [12] [29]:

- Calcular la similitud del nuevo documento con todos los documentos de la colección.
- Determinar los k documentos más similares al nuevo documento.
- Asignar la clase del nuevo documento, que corresponde a aquella que ocurre más frecuentemente en los k documentos seleccionados.

K-nn es uno de los clasificadores de textos más utilizados y que ha mostrado uno de los mejores desempeños [84]. Su única desventaja es el relativo alto costo computacional debido que para decidir si un documento d pertenece a una categoría c, se debe revisar si los k documentos de entrenamiento más similares a d pertenecen a la categoría c, y esto se debe realizar para cada documento del conjunto de testeo [29].

Para utilizar el algoritmo se debe elegir el valor de k, el cual es probable que se pueda elegir a priori o bien por inspección de casos.

Para métodos tradicionales de k-nn usualmente no se tiene un muy buen desempeño para clasificación de documentos. El grado de desempeño puede mejorar notoriamente seleccionando una medida apropiada de similitud [84], por ejemplo similitud de coseno que ha sido ampliamente utilizada y con excelente resultados para text mining.

2.3.7.3. Evaluación de clasificación supervisada

Existen dos estrategias principales para realizar la validación de clasificadores [12]. La primera estrategia es separar los datos en un conjunto de entrenamiento y un conjunto de testeo, donde el conjunto de entrenamiento es utilizado para construir el clasificador que luego es usado para predecir la clasificación de las instancias del conjunto de testeo, con el objetivo de finalmente medir el desempeño del modelo basado en las instancias correctamente clasificadas. La segunda estrategia es usar validación cruzada de K iteraciones, en la que las instancias son divididas en K partes iguales y se realizan K iteraciones, en cada una de dichas iteraciones se toma una de las K partes como conjunto de testeo y las K-1 partes restantes como conjunto de entrenamiento.

Una vez realizada la clasificación, para cada clase C_k se puede construir la siguiente matriz de confusión:

		Clase Predicha	
		C_k	**Distinta de C_k**
Clase Real	**C_k**	TP	FN
	Distinta de C_k	FP	TN

Cuadro 2: Matriz de confusion para la categoria C_k
Fuente: [12]

TP es el número de instancias perteneciente a la clase C_k clasificadas correctamente como clase C_k, FP el el número de instancias pertenecientes a una clase distinta de C_k pero clasificadas como clase C_k (error tipo 1), FN es el número de instancias pertenecientes a la clase C_k pero clasificadas como una clase distinta de C_k (error tipo 2) y TN es el número de instancias pertenecientes a una clase distinta de C_k clasificadas correctamente como una clase distinta de C_k. Para un clasificador perfecto FP y FN deberían ser ambos cero. Dado un conjunto de testeo se puede definir

distintas medidas de desempeño para un clasificador, las más importantes son las siguientes:

Medida	Fórmula	Descripción
Recall	TP/(TP+FN)	Proporción de instancias de la clase C_k que son correctamente clasificadas como C_k.
Precisión	TP/(TP+FP)	Proporción de instancias clasificadas como C_k que son realmente de la clase C_k.
F-measure	(2 x precisión x recall) / (precisión+recall)	Una medida que combina precisión y recall
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Proporción de instancias que son correctamente clasificadas.
Error Rate	(FP+FN)/(TP+TN+FP+FN)	Proporción de instancias que son incorrectamente clasificadas.

Cuadro 3: Medidas de desempeño para modelos de aprendizaje supervisado
Fuente: [12]

Para clasificación con múltiples categorías las medidas de desempeño globales del clasificador pueden ser calculadas de dos formas: micro- averaging o macro-averaging [64] [85]. Sea M la cantidad de categorías, F-measure (micro-averaged) es calculado globalmente para todas las categorías, en la que precisión y recall son obtenidas sumando los valores individuales de sus componentes:

$$\boldsymbol{Precisión_{micro}} = \frac{\sum_{i=1}^{M} \boldsymbol{TP_i}}{\sum_{i=1}^{M} (\boldsymbol{TP_i} + \boldsymbol{FP_i})} \qquad \boldsymbol{Recall_{micro}} = \frac{\sum_{i=1}^{M} \boldsymbol{TP_i}}{\sum_{i=1}^{M} (\boldsymbol{TP_i} + \boldsymbol{FN_i})}$$

$$\boldsymbol{F - measure_{micro}} = \frac{2 \times \boldsymbol{Precisión_{micro}} \times \boldsymbol{Recall_{micro}}}{\boldsymbol{Precisión_{micro}} + \boldsymbol{Recall_{micro}}}$$

Por otro lado, para las medidas macro-averaged se calculan los valores localmente para cada categoría primero y luego se toma el promedio de todas las categorías:

$$\boldsymbol{Precisión_{macro}} = \frac{1}{M} \sum_{i=1}^{M} (Precisión_i) = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{\boldsymbol{TP_i}}{\boldsymbol{TP_i} + \boldsymbol{FP_i}} \right)$$

$$\boldsymbol{Recall_{macro}} = \frac{1}{M} \sum_{i=1}^{M} (Recall_i) = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{\boldsymbol{TP_i}}{\boldsymbol{TP_i} + \boldsymbol{FN_i}} \right)$$

$$\boldsymbol{F - measure_{macro}} = \frac{1}{M} \sum_{i=1}^{M} (F - measure_i) = \frac{1}{M} \sum_{i=1}^{M} \frac{2 \times \boldsymbol{Precisión_i} \times \boldsymbol{Recall_i}}{\boldsymbol{Precisión_i} + \boldsymbol{Recall_i}}$$

Las medidas de desempeño obtenidas bajo micro-averaged dan igual ponderación a cada uno de los documentos, mientras que las medidas macro-averaged dan igual ponderación a cada categoría, independientemente de su frecuencia [64]. Los valores más altos para cada medida corresponden a una calidad de clasificación más alta.

2.3.8. Clustering de documentos

Clustering es una técnica de aprendizaje automático que tiene por objetivo agrupar un conjunto de objetos en subconjuntos o clusters [46]. Un cluster consiste en un grupo de objetos en que sus miembros son muy similares entre ellos y substancialmente distintos a los objetos de los otros grupos [38] [39]. Todos los problemas de clustering son, en esencia, problemas de optimización [29], en los que se maximiza o minimiza una función de calidad, generalmente en términos de similitud entre los objetos, sujeta a un conjunto de restricciones.

Un algoritmo de clustering minimiza la distancia intra cluster y maximiza la distancia inter cluster, usando una apropiada medida de distancia entre los objetos [46], es decir, se agrupan los objetos similares y se separan aquellos que no lo sean.

A diferencia del problema de clasificación, en el problema de clustering no hay clases predefinidas, dado que no se tiene la información a priori sobre la clase a la que pertenece un objeto, aunque cada cluster que se forma puede ser visto como una clase de objetos [39]. Clustering es una de las formas más comunes de aprendizaje no supervisado [46],

El análisis de clusters permite contribuir a la comprensión de la información cuando la cantidad de datos es muy grande y su procesamiento muy demandante, entregando una estructura adicional a la muestra de datos a través del descubrimiento de distribuciones y correlaciones entre los conjuntos de datos. También permite inferir algunas hipótesis referentes a los datos o validar hipótesis específicas [38] [84].

Dentro de las categorías de clustering más utilizados destacan dos tipos [45]: clustering particional y clustering jerárquico. En clustering particional se intenta directamente descomponer el conjunto de objetos en un conjunto de clusters disjuntos, optimizando algún criterio predefinido o función objetivo en un procedimiento iterativo. En clustering jerárquico se procede sucesivamente juntando clusters de un miembro cada uno hasta construir un solo cluster con todos los objetos (método aglomerativo) o viceversa (métodos divisivos) y el resultado es un árbol de clusters (dendrograma) en el cual se muestra como los clusters están relacionados. Para clustering de grandes conjuntos de documentos de textos los algoritmos de clustering particionales resultan más adecuados, debido a sus requerimientos computacionales relativamente bajos [75].

Considerando el constante incremento de contenido de textos (contenidos disponibles en internet, bibliotecas digitales e información personal digitalizada) la aplicación de clustering de documentos resulta muy útil para tareas como [23] [46]: encontrar documentos similares, organizar y buscar eficientemente en grandes colecciones de documentos, detectar contenidos duplicados (detección de plagio), etiquetar colecciones de documentos, entre otros. Para obtener resultados útiles y

eficientes a partir del clustering de documentos se debe realizar primeramente una apropiada selección de atributos, definir el tipo de medida de similitud y especificar el algoritmo de clustering a utilizar.

En clustering de documentos se asigna cada documento a un cluster, agrupando los documentos similares dentro de un mismo cluster y separándolos de los otros clusters formados por documentos diferentes [35]. En otras palabras, los documentos en un cluster comparten una temática y los documentos en diferentes clusters representan distintas temáticas [59]. Para este tipo de clustering la medida de similitud asume un rol muy importante [44], dado que los documentos no poseen una clase predefinida por lo que la similitud refleja el grado de cercanía o separación de los objetos y define la asignación de clases.

Los clusters de documentos formados tienen un significado implícito en su agrupamiento, por lo que una correcta caracterización de un cluster mediante ciertas palabras claves puede entregar un significado útil para la interpretación de cada cluster [81] [84]. Los resultados de la aplicación de clustering deben ser evaluados por algún método definido previamente e interpretados correctamente basados en alguna otra evidencia experimental [38].

Los mayores inconvenientes que puede tener el clustering de documentos son la alta dimensionalidad del espacio de atributos que afecta su eficiencia y los atributos redundantes o irrelevantes que pueden alterar los resultados [59].

2.3.8.1. Algoritmo K-means

K-means es un proceso iterativo de clustering particional [44] y además es un tipo de clustering exclusivo en el que cada objeto es asociado a un único cluster [12]. K-means está basado en la idea de que un punto central puede representar un cluster [75].

El algoritmo K-means particiona una colección de vectores $\{x_1, x_2,\ldots, x_n\}$ en un conjunto de clusters $\{C_1, C_2,\ldots, C_k\}$, para lo cual se necesita k centroides para inicializar el algoritmo, los que pueden ser proporcionados externamente o elegidos aleatoriamente entre los vectores [29]. La solución generada por el clustering son óptimos locales [44] dados por el conjunto de datos, el valor de k y los centroides iniciales. Una vez finalizado el algoritmo cada vector es asignado a un cluster específico y cada cluster puede ser visto como una clase representativa, por lo que cualquier método de aprendizaje para clasificación puede ser aplicado [84].

El objetivo del algoritmo k-means estándar [46] [89] es minimizar el error cuadrático medio de todos los clusters, en otras palabras, minimizar la varianza intra cluster total.

El error cuadrático medio de un cluster se define como la media de los cuadrados de las distancias de los objetos de un cluster al centroide del cluster, donde el centroide de un cluster C, μ_c, está dado por:

$$\boldsymbol{\mu_C} = \frac{\mathbf{1}}{|\boldsymbol{C}|}\sum_{\vec{\boldsymbol{x}}\in\boldsymbol{C}}\vec{\boldsymbol{x}}$$

K-means es un algoritmo clásico de clustering siendo uno de los más populares debido a su simplicidad y eficiencia [29]. En general k-means converge después de relativamente pocas iteraciones, pero a medida de que el número de clusters, k, se hace más grande la eficiencia del algoritmo disminuye [84].

Uno de los principales inconvenientes que puede presentar k-means es la determinación del número de clusters, k, que se formarán a partir de los datos, dado que el algoritmo no entrega información sobre el mejor valor de k. Seleccionar un mal valor de k generará un resultado sub-óptimo y que podría no responder a las metas originales de la investigación [29]. En general el valor de k corresponde a un número entero relativamente pequeño [12]. Existen varios métodos para tratar de encontrar un buen valor de k [84], por ejemplo realizar varias iteraciones del clustering con diferentes valores de k aleatorios y elegir el mejor k de acuerdo a alguna función de calidad, así como también si se tiene algún conocimiento previo sobre la naturaleza de los documentos se podría inferir algún número razonable de categorías a utilizar.

El algoritmo k-means estándar consiste en los siguientes pasos [12] [29] [75]:

- Inicialización:
  - Se elige un valor de k.
  - Se seleccionan k documentos de forma aleatoria como los centroides iniciales.
  - Los restantes documentos son asignados a los clusters que tengan su centroide más cercano o sea más similar.

- Iteración:
  - El centroide de cada cluster es recalculado basado en los actuales miembros del cluster.
  - Cada documento es reasignado al cluster que tenga su centroide más cercano o sea más similar.

- Detención:
  - Como criterio de detención se pueden usar distintas condiciones, por ejemplo una vez que los centroides no cambien o bien definir un máximo de iteraciones.

Aplicado k-means a documentos, estos comienzan todos agrupados y luego son distribuidos en grupos más pequeños de documentos similares, realizándose varias iteraciones de este proceso hasta alcanzar un criterio detención. El algoritmo k-means

ha sido utilizado extensamente con documentos de textos, presentando una alta eficiencia [75] [84].

Una variación del método estándar de k-means euclidiano es el algoritmo k-means esférico (k-means utilizando similitud de coseno), que corresponde a un popular método para clustering de textos de alta dimensionalidad que tiene ventajas desde una perspectiva computacional siendo uno de los métodos más eficientes [23] [89].

2.3.8.2. Evaluación de clustering

Una de las tareas más relevantes y complejas de la aplicación de algoritmos de clustering es la evaluación de los resultados conocida como validación de clusters [38] [69]. Dado que clustering es un proceso no supervisado, las medidas de desempeño usadas para el caso clasificación no se pueden utilizar [50].

Los resultados obtenidos a partir de un algoritmo de clustering dependen de los parámetros de entrada, como el número óptimo de clusters que se ajustan al conjunto de datos (por ejemplo en k-means). Para evaluar y seleccionar un esquema de clustering dos medidas resultan fundamentales [49]: compacidad y separación. Se espera que los miembros dentro de cada cluster estén tan cerca como sea posible y que los clusters entre sí estén ampliamente separados.

Un enfoque común para la evaluación de la calidad de los resultados de clustering es usar índices de validación de clusters [69] [87] que tienen por objetivo encontrar un conjunto de clusters que se ajuste a una partición natural de los datos, usualmente definidos combinando compacidad y separabilidad.

Existen tres diferentes enfoques para estudiar la validez del resultado de un algoritmo de clustering [14] [38] [49] [50] [69]: criterios externos, criterios internos, criterios relativos.

Los criterios externos evalúan los resultados basados en una estructura de información preespecificada, la cual es impuesta en el conjunto de datos, utilizando la intuición específica del usuario acerca de la estructura de clustering del conjunto de datos. El principal objetivo es determinar la calidad del algoritmo de clustering para reconocer grupos existentes.

Los criterios internos estudian los resultados basados en las características propias de los clusters, sin información adicional acerca de los datos o repeticiones del proceso de clustering. El objetivo primordial es determinar la calidad del algoritmo para generar interesantes particiones. Dentro de los criterios internos, el índice Davies-Bouldin (DB) [22] es uno de los más utilizados y con menor complejidad computacional. Este índice

tiene por objetivo identificar el conjunto de clusters que son compactos y bien separados a través de una función de la proporción entre la suma de la dispersión intra-cluster y la separación inter-cluster [53] [57]. El índice Davies-Bouldin está definido como

$$\boldsymbol{DB} = \frac{\mathbf{1}}{\boldsymbol{c}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{c}} \left(\max_{\boldsymbol{j}=\mathbf{1},\ldots,\boldsymbol{c};\, \boldsymbol{i}\neq\boldsymbol{j}} \left(\frac{\boldsymbol{d}(\boldsymbol{X_i}) + \boldsymbol{d}(\boldsymbol{X_j})}{\boldsymbol{d}(\boldsymbol{c_i}, \boldsymbol{c_j})} \right) \right)$$

donde c denota el número de clusters, $\boldsymbol{d}(\boldsymbol{X_i})$ es la dispersión intra-cluster del cluster i y $\boldsymbol{d}(\boldsymbol{c_i}, \boldsymbol{c_j})$ es la separación entre los centroides del cluster i y j. Los valores más pequeños del índice DB representan un mejor resultado del clustering, por lo que el número de cluster que minimiza el índice DB es el número óptimo de cluster [67] [87], indicando un esquema en que los clusters están de forma compacta y los centros de los clusters están bien separados entre sí.

Los criterios relativos analizan los resultados basados en la comparación de esquemas de clustering ejecutados varias iteraciones pero con diferentes valores de los parámetros de entrada o distintos subconjuntos de datos. Este enfoque no utiliza información adicional de los datos. El principal objetivo es determinar la calidad del algoritmo para generar grupos significativos.

En la mayoría de los casos prácticos no se tiene un conocimiento previo sobre el conjunto de datos, por lo que aplicar criterios externos de validación de clusters resulta muy complejo. A partir de los experimentos realizados en [69], los criterios internos resultaron ser más precisos que los criterios externos.

Una vez que el proceso de validación de clusters ha sido realizado, es posible que sea útil evaluar también la capacidad predictiva de la asignación de clusters [84]. Para estimar la capacidad predictiva se puede utilizar cualquier método de aprendizaje supervisado estándar y calcular las medidas definidas para evaluación de clasificación (por ejemplo accuracy, recall y/o precisión), aunque debe considerarse que la evaluación del desempeño obtenida será optimista debido a que en el proceso de clustering se analizaron todos los datos (tanto el conjunto de entrenamiento como de testeo).

2.3.9. Herramientas complementarias

2.3.9.1. Visualización de textos

Herramientas de visualización para text mining pueden tener una crucial importancia para facilitar el descubrimiento de conocimiento, ya que entrega un

panorama general de una gran cantidad de datos [11]. Un tag cloud es un simple y ampliamente usado modelo de interfaz visual que representa el resumen de un documento o colección de ellos a través de una lista de palabras, generalmente presentada en orden alfabético, cuya importancia relativa de cada palabra es caracterizada por el tamaño de la letra, el color y/o la ubicación. El criterio para la selección de las palabras comúnmente es la frecuencia de ocurrencia de la palabra, por lo que puede ser considerado similar a un histograma. Un tag cloud logra sintetizar grandes volúmenes de documentos de forma comprensible.

Métodos no supervisados de text mining, como clustering, requieren un intenso trabajo para una correcta interpretación de los resultados para conseguir conclusiones útiles, pero herramientas visuales de post-procesamiento pueden facilitar esta tarea [11]. A través de diversos estudios [40] [41] [65] se concluye que la formación de tag clouds puede tener un valioso papel para confirmar la información generada con técnicas de minería de textos, particularmente cuando se trabaja con clustering, por ejemplo la aplicación de tag clouds a cada una de las particiones generadas permite confirmar si los clusters son coherentes y significativos, donde cada clusters debería estar caracterizado por varias palabras representativas (con alta frecuencia).

2.3.9.2. Reglas de asociación

Las reglas de asociación, introducidas en [3], son una forma simple y útil de descubrir conocimiento de forma eficiente en grandes volúmenes de datos, entregando información de las reglas generadas entre los datos. Sean dos conjuntos disjuntos, A y C, una regla es una implicación $A \rightarrow C$, donde la parte izquierda de la regla se denomina antecedente y la parte derecha consecuente.

Entre los criterios que permiten determinar un nivel de validez de una regla se encuentra el soporte y la confianza de una regla [19] [88], dos de los criterios más usados. Soporte se define como la fracción de registros que contienen los elementos de los conjuntos de antecedente y consecuente con respecto al total de registros, lo cual entrega una significancia estadística de una regla de asociación. Por otro lado, confianza corresponde a la fracción de registros que contienen los elementos de los conjuntos de antecedente y consecuente con respecto al total de registros que contienen los elementos del antecedente, lo cual puede ser visto como una forma de medir la fuerza de las reglas de asociación.

En el contexto de text mining, una regla de asociación dado un conjunto de términos $\{t_1,\ldots, t_n\}$, es de la forma [19]: $\{t_1,\ldots, t_k\} \rightarrow \{t_{k+1},\ldots, t_n\}$, cuyo interpretación es una relación de que la ocurrencia de los términos $\{t_1,\ldots, t_k\}$ implican la ocurrencia de los $\{t_{k+1},\ldots, t_n\}$ en un documento. El soporte de esta regla está definida como la proporción

de documentos que contienen ambos conjuntos de términos, $\{t_1,\ldots, t_k\}$ y $\{t_{k+1},\ldots, t_n\}$, con respecto al total de documentos. La confianza de esta regla está definida como la probabilidad de que el conjunto de términos $\{t_{k+1},\ldots, t_n\}$ aparezca en el documento dado que aparece el conjunto de términos $\{t_1,\ldots, t_k\}$ y mide la proporción de documentos que cumplen la regla.

Uno de los principales problemas con esta técnica es la gran cantidad de reglas que se generan, algunas de las cuales suelen ser redundantes, por lo que determinar aquellas más interesantes puede convertirse en una difícil tarea, especialmente cuando el interés está puesto en el comportamiento de pequeños subgrupos del conjunto original de datos [47]. Para llevar a cabo el proceso de extracción de reglas de asociaciones útiles es necesario determinar entonces criterios de calidad de cada regla individualmente [12], por lo que se definen valores mínimos de soporte y confianza que debe cumplir una regla de asociación dentro de un conjunto de datos [19] [66] [88]. En un primer paso de buscan todos los conjuntos de elementos que tengan un soporte mayor al umbral mínimo definido y, a partir de los conjuntos encontrados, se seleccionan solo aquellos conjuntos cuya confianza supere un umbral mínimo.

En [66] y [70] la generación de reglas de asociación se utiliza de forma combinada con algoritmos de clustering. Dentro de cada partición o cluster, generado a partir de algún algoritmo de clustering, se extraen pequeños y confiables conjuntos de reglas de asociación, usando los principales atributos que determinan cada cluster. De esta forma se logra mejorar la caracterización de cada una de las particiones.

2.4. Crime mining

Una de las tareas primordiales de las fuerzas policiales es la prevención del delito, para lo cual el análisis de registros policiales resulta clave, dado que permite el diseño de políticas de prevención oportunas y efectivas [79].

El análisis especializado de crímenes es realizado, en general, por especialistas con formación en investigación criminal, sicología e informática [63]. Los principales métodos de investigación utilizados en análisis criminal son el análisis de tendencias y patrones delictivos junto con la identificación de vínculos entre tipos de crímenes y grupos de delincuentes. Este tipo de análisis manual es un proceso que consume mucho tiempo y es propenso a errores, y está basado principalmente en herramientas de estadística descriptiva clásica [79]. En general, las cantidades de datos disponibles sobre delitos son enormes y muy poco utilizados, por ejemplo las colecciones de reportes policiales [6], cuyo análisis puede revelar interesante información útil para la prevención del delito.

Dentro de las principales dificultades del análisis de delitos es la gran cantidad de datos involucrados, la no estructura de los datos (en general en formato de texto) y que la estadística descriptiva clásica no responde las necesidades de un problema que requiere un tratamiento estadístico más complejo.

Data mining se transforma en una poderosa herramienta para el análisis de datos, estructurados y no estructurados, que permite a los investigadores criminales, que pueden ser no expertos en análisis de datos, explorar grandes bases de datos de forma rápida y eficiente [18]. La aplicación de minería de datos a la información criminal entrega un alto valor agregado por la generación de nuevo conocimiento considerando la gran cantidad de información que no es aprovechada totalmente [8]. La integración de nuevas tecnologías como la minería de textos [6] [60] [63] permite a las policías descubrir interesantes patrones criminales que pueden ayudar a mejorar la tasa de resolución de crímenes, diseñar planes de prevención efectivos, incrementar la eficiencia operativa reasignando las tareas de los analistas, reducir los tiempos de análisis y disminuir los potenciales errores.

Entre las principales técnicas utilizadas en crime mining se encuentran [17] [18] [60] [80]:

- Clasificación: encuentra propiedades comunes entre diferentes entidades criminales y las organiza en clases predefinidas, estas técnicas han sido usadas para identificar fuentes de email spamming y para predecir tendencias criminales.

- Clustering: se agrupan datos dentro de una clase con características similares, útiles para la predicción y asociación de crímenes, por ejemplo para identificar sospechosos que comenten varios crímenes de forma similar. Se focaliza en la detección de grupos de crímenes que ocurren en una zona geográfica determinada. Cada grupo de crímenes formado posee una lista de características particulares que sirven para mejorar la vigilancia táctica de la policía.

- Otras técnicas: rule induction permite buscar posibles explicaciones de una tendencia o comportamiento, por ejemplo encontrar posibles causas del incremento de robos; la extracción de entidades permite identificar patrones específicos desde los datos y ha sido utilizado para detectar automáticamente personas y vehículos desde los reportes policiales; las reglas de asociaciones permiten descubrir elementos frecuentes y presentar los patrones como reglas, y ha sido aplicada en la detección de intrusiones de red; y la detección de outlier identifica los datos notoriamente distintos del conjunto de datos y puede ser utilizado en la detección de fraudes.

El crimen se caracteriza por ocurrir en un lugar y momento determinado, por lo que un análisis espacial y temporal puede ser bastante apropiado. Los sistemas de información geográfica (SIG) han permitido situar distintos fenómenos sociales como los delitos [8] [63] basado en la visualización de las demarcaciones zonales y en información propia del hecho delictual: tipo de delito, fecha, lugar, cantidad de víctimas,

etc. De esta forma se pueden detectar las zonas más propensas a sufrir ciertos tipos de crímenes [6] [80], mejorando también la productividad de las labores del personal de orden [60].

En [18] se establece una lista de 8 categorías de crímenes (basadas en la clasificación de diferentes fuerzas de orden y especialistas en el tema criminal), que afectan a la población y de los cuales las policías comúnmente mantienen gran cantidad de datos disponibles para realizar distintos análisis.

Tipo de Crimen	Descripción
Violaciones de Tránsito	Conducción peligrosa, daño a la propiedad o lesiones a las personas durante una colisión y conducción bajo efectos del alcohol o drogas.
Delitos Sexuales	Abuso sexual, violación, prostitución y acoso de menores.
Robos	Robos con fuerza, hurto y robo de vehículos.
Fraudes	Lavado de dinero, falsificación, corrupción y soborno.
Incendios Intencionales	Intencionalmente prender fuego para causar daño a la propiedad.
Delitos de Drogas	Posesión, distribución y ventas de drogas ilegales.
Crímenes Violentos	Homicidios y robos con arma de fuego.
Cibercrimen	Fraudes a través de internet (fraudes con tarjetas de crédito), intrusión de redes y hacking, propagación de virus, ciberpiratería y ciberterrorismo.

Cuadro 4: Categorías de Crímenes con gran cantidad de datos para análisis
Fuente: [18]

2.4.1. Sistema de información geográfica

Los distintos delitos pueden ser estudiados considerando la posible influencia que tengan factores geográficos y sociales en el nivel de ocurrencia y características de un tipo de delito, cuyo principal interés es la identificación de patrones de los delitos y analizar si existen ciertas tendencias que permitan agrupar los casos. La mayoría de las agencias policiales utiliza sistemas de información para el análisis de investigación criminal, usando los datos que obtienen a través de los reportes policiales (con el lugar y fecha del delito) [4] [15].

Los sistemas de información geográfica (SIG) [42] [83] permiten realizar un seguimiento de un amplio rango de fenómenos sociales, en particular se ha utilizado como una herramienta para el análisis de ocurrencia de delitos, basado en el supuesto de que los incidentes ocurren de alguna forma predecible. En un SIG se usan herramientas para la captura, el almacenamiento, el análisis y la visualización de información georreferenciada. Un SIG cuenta con una base de datos geográfica que

enlaza con información descriptiva, lo cual permite analizar distintos factores influyentes de un hecho criminal. El conocimiento derivado de un SIG puede ser tan simple como un resumen geográfico que señale las locaciones de todos los eventos que ocurren en un período de tiempo determinado en una zona específica.

El principal objetivo de la utilización de un sistema de información geográfica es mejorar el entendimiento de las relaciones entre los distintos atributos que caracterizan un delito, en particular investigar la relación entre un delito y la zona en que este ocurre. Lo anterior se traduce en la identificación de patrones para [4] [15] [83]: la detección de zonas con altos niveles de ocurrencia de un tipo de delito (conocido como hotspots), la exploración del vínculo entre la actividad criminal y factores ambientales y/o socio-económicos, el estudio del movimiento de los delitos para predecir la locación de futuros delitos y determinar si la tendencia en la ocurrencia de un tipo específico de delito está o no relacionada con una localización particular. La utilización de un SIG permite al personal policial realizar un plan efectivo ante emergencias, delimitar zonas prioritarias de vigilancia, analizar eventos pasados y predecir posibles eventos futuros. Además puede ser útil durante el diseño de estrategias de prevención y para la evaluación de los programas policiales aplicados en el pasado, así como para una eficiente y efectiva asignación de tareas de los recursos policiales.

2.5. News mining

Las noticias corresponden a un tipo de información pública que presenta características muy particulares, diferentes de otros tipos de textos [72]. Proveen grandes y recurrentes cantidades de recursos de información [62], tanto de eventos actuales como históricos [52] y cuyo propósito es provocar un impacto en el receptor [9]. Las noticias reflejan el punto de vista de una sociedad, grupo o individuo sobre alguna temática de interés de forma casi instantánea una vez que los eventos noticiosos ocurren, escritos en general en un formato y lenguaje de reporte periodístico [9] [58].

La mayoría de los periódicos de noticias disponen su contenido en sus sitios web [62], así como también han surgido cada vez más periódicos de noticias online. Debido a lo anterior, nuevas herramientas surgen para acceder a estas nuevas tecnologías de forma más eficiente [1] [62] [72], por ejemplo los RSS feeds que permiten acceder al contenido de forma unificada, personalizada y dinámica. Los RSS feeds logran manejar un creciente número de fuentes de información fiable con una rápida actualización de contenidos.

Los servicios de noticias online entregan cientos de noticias diariamente por lo que un proceso de etiquetado de cada noticia con alguna temática puede responder a las necesidades particulares de un lector cuando desea seguir regularmente una temática

específica. Este proceso de etiquetamiento puede realizarse de forma manual pero es demasiado costoso en tiempos de ejecución e inconsistente, dado que dos personas pueden tener distintas puntos de vista para una misma noticia. El problema debe ser resuelto basado en un método automatizado para la asignación de temáticas de los artículos noticiosos [84], por lo que las técnicas de minería de textos toman mucha importancia. La solución a este problema involucra clasificación de documentos o bien clustering de documentos, aunque en general la información de las temáticas de las noticias no se encuentra disponible en las diversas fuentes, por lo que la alternativa de clustering toma mayor relevancia [62] [72].

El análisis de colecciones de noticias permite estudiar distintos intereses y comportamientos de una sociedad (y constatar si estos cambian o se mantienen en el tiempo) [56], dado que es esperable que las temáticas tratadas en las noticias tengan una alta correlación con los intereses de una sociedad [58] En general, este tipo de análisis en las noticias se realiza sobre los reportes, dejando fuera las columnas de opinión o comentarios [9]. Dentro de los principales análisis que se pueden realizar a partir del estudio de noticias se encuentran [52] [72]: caracterización de noticias, predicción de temáticas (supervisada o no supervisada), detección de tendencias, búsqueda de contenido, análisis espacial y temporal de una temática, personalización según intereses del lector, detección automática de eventos e identificación de entidades (personas, lugares, empresas, etc.). En general, cualquiera de las distintos tipos de análisis automatizados de noticias resulta ser un problema complejo [62].

Uno de los ámbitos prácticos en que se ha utilizado news mining es en el mercado de las acciones [26] [61], en el cual se ha reconocido la influencia del contenido de las noticias económicas como factor relevante en la tendencia de los precios. De esta forma las noticias económicas permiten apoyar los modelos basados en los precios históricos del mercado para lograr predicciones más precisas, identificando las noticias con impacto positivo o negativo en los precios.

3. Metodología de investigación

La metodología de investigación adoptada en este trabajo está basado en la serie de etapas que caracterizan el modelo de proceso CRISP-DM: comprensión del ámbito de aplicación, comprensión de los datos, preparación de los datos, modelado, evaluación y desarrollo.

3.1. Comprensión del ámbito de aplicación del estudio

El contexto del estudio es el análisis de las noticias policiales nacionales tomando en consideración la relevante influencia que tendrían dentro de la sociedad. Se pretende identificar y caracterizar los distintos tipos de delitos relatados por los medios de prensa, además de revisar la relación entre los niveles reales de criminalidad con lo informado por los medios. Ya en la sección 2.1. se señalaron diversas teorías sobre lo que una distorsión de los medios de prensa puede ocasionar en los niveles de inseguridad de las personas, de allí que un estudio práctico y apoyado en tecnologías eficientes de procesamiento de datos resulta interesante.

Se revisa la factibilidad de obtener los recursos necesarios para llevar a cabo este trabajo, como lo son la disponibilidad de un volumen considerable de noticias y estadísticas oficiales de los casos policiales. Por otro lado, se entiende que el estudio presenta ciertas restricciones, dado que se analiza un subconjunto de noticias del total disponible durante un período de tiempo acotado.

3.2. Comprensión de los datos

En esta etapa se seleccionan las noticias que serán parte del estudio. Se escogen 4 medios de prensa distintos que ofrecen sus contenidos a través del formato RSS. Los medios de prensa distribuyen su contenido bajo distintas secciones de interés, pero no todos poseen la sección “policial” identificada para sus usuarios, por lo que se utiliza en su defecto, en una primera instancia, la sección “nacional” para formar la base de datos. Cada noticia es almacenada con su titular, descripción (contenido), fecha de publicación, medio de prensa de origen y región de ocurrencia (esta última es directamente extraída a partir del contenido de la noticia y cada noticia puede estar asociada a una o más regiones). Se verifica la calidad de los datos a través de la revisión de la existencia de campos vacíos y se eliminan los registros repetidos.

A continuación se presenta la distribución de las noticias seleccionadas inicialmente[3]:

- Fuente Alfa, sección policial: 6.464 noticias (julio-diciembre 2011).
- Fuente Beta, sección policial: 2.374 noticias (julio-diciembre 2011).
- Fuente Gamma, sección nacional: 6.286 noticias (julio-diciembre 2011).
- Fuente Delta, sección nacional: 5.937 noticias (julio-diciembre 2011).
- Fuente Beta, sección nacional: 2.665 noticias (noviembre-diciembre 2011).

De esta forma, la base de datos inicial contiene 23.726 noticias distintas a ser procesadas.

3.3. Preparación de los datos

Dentro de esta etapa se desarrolla: la lista de stop-words, el stemming y la selección de atributos.

La lista de stop-words se forma con palabras con dos o menos caracteres o bien con más de veinte, palabras propias de un medio web (www, http, xml, etc.), preposiciones, artículos, adverbios, pronombres, conjunciones, nombres propios de personas, regiones y localidades, apellidos, meses, días, números y conjugaciones de los verbos: ser, estar, haber, hacer, ir y tener. Toda la lista es revisada antes de ser aplicada, debido a la posibilidad de la existencia de alguna palabra que pueda ser relevante para el contexto tratado.

Se utiliza un método de stemming ajustado a la base de datos y para ello se elabora una lista de conversiones cumpliendo las siguientes reglas:

- Plural a singular.
- Femenino a masculino.
- Palabras con sufijo –miento, –ción y -sión a verbo infinitivo.
- Eliminación del sufijo –mente.
- Conjugaciones verbales a verbos infinitivos.

Dicha lista de conversiones se diseña basada en la aparición de las palabras involucradas dentro de la base de datos, por lo que el stemming puede responder eficientemente a las necesidades de reducción del número de atributos en el contexto de esta investigación particular, pero probablemente no será muy útil para otra aplicación. Bajo este mecanismo la base de datos formada por 23.726 noticias pasa de tener 61.535 palabras distintas a 32.742.

[3] Se utilizarán las noticias de 4 medios de prensa distintos, los cuales serán presentados con alias con el fin de evitar cualquier tipo de juicio particular a un medio de prensa a partir de los resultados, lo cual no forma parte de la intención del presente trabajo.

La selección de atributos se realiza de forma independiente para las bases de datos de cada una de las fuentes noticiosas. Como filtro básico de selección de atributos se utiliza que las palabras deben aparecer al menos en un 0,5% de los documentos del conjunto de documentos para fines de clasificación y un 0,1% de los documentos para fines de clustering (que corresponde a la medida de document frequency). Luego se utilizará para clasificación las medidas chi-cuadrado e information gain, descritos en la sección 2.3.6.1, seleccionando las palabras con los más altos valores para ambas medidas simultáneamente. Para clustering se utiliza TVQ, TV y TC, descritos en la sección 2.3.6.2., seleccionando por cada medida las palabras con los más altos valores.

Las noticias son representadas como vectores binarios (presencia o ausencia de una palabra), no considerando ningún tipo de esquema de ponderación de un término. Esta representación se usa principalmente porque la agrupación de temáticas de noticias se realiza basada en la co-ocurrencia de las palabras, siendo menos relevante la ponderación individual que se le pueda entregar a una palabra.

3.4. Modelado[4]

En una primera etapa se requiere formar una base de datos con únicamente noticias de índole policial, por lo que se debe encontrar un modelo de clasificación que permita distinguir, dentro de las noticias de la categoría nacional, las noticias policiales de las que no lo son. Luego de formada la base de datos únicamente con noticias policiales se puede iniciar el proceso de identificación de temáticas policiales dentro de las noticias.

3.4.1. Aplicación de métodos de aprendizaje supervisados

Para realizar el estudio sobre noticias policiales, se requiere primeramente discriminar aquellas noticias policiales de las no policiales, en aquellos medios de prensa que no disponen de la sección policial. Para lo anterior se construye un modelo de clasificación (Policial/No Policial) a partir de la base de datos formada por las noticias de la sección nacional de la fuente Beta. En dicha base de datos se dispone de las etiquetas de clase Policial, Política y Actualidad, mientras que para el modelo a construir se utilizan las etiquetas Policial y No Policial (esta última compuesta por las etiquetas Política y Actualidad).

[4] Todo el proceso que involucra la construcción y evaluación de modelos de aprendizaje automático se utiliza el programa informático RapidMiner, que corresponde a un sistema de código abierto para análisis y minería de datos (http://rapid-i.com/)

Se estudiarán dos modelos de clasificación distintos: Naive Bayes y K-nn, seleccionando el modelo con mejor desempeño basado en el método de evaluación k-fold cross validation. El modelo seleccionado es aplicado luego a las bases de datos formadas por las noticias de las secciones nacionales de las fuentes Gamma y Delta. De esta forma la nueva base de datos queda constituida solo por noticias de carácter policial.

3.4.2. Aplicación de métodos de aprendizaje no supervisados

Las bases de datos con las noticias policiales de cada medio de prensa se utilizan independientemente para detectar las principales temáticas tratadas. El algoritmo de clustering elegido para esta tarea es K-means. Realizada la etapa de pre-procesamiento, incluyendo la selección de atributos, el número de palabras que pueden caracterizar una temática ha descendido notoriamente, por lo que la tarea se puede desarrollar de forma más eficiente.

Debido a que se desconoce el número de posibles temáticas tratadas, se deben realizar varias iteraciones del algoritmo a cada una de las bases de datos, con el fin de detectar posibles candidatos de temáticas o bien grupos de palabras que puedan estar generando distorsión. Hasta este punto no se han eliminado las palabras que presentan una alta frecuencia (ya que no era posible suponer aún si una palabra frecuente sería o no útil para discriminar una temática), por lo que se diseña un esquema básico para una última selección de atributos: a partir de un esquema de clustering con 10 particiones, se eliminan las palabras que cumplen algunas de las siguiente dos condiciones:

- Palabras que aparecen en más de un 10% de los documentos que forman un cluster, para más de 7 clusters.
- Palabras que aparecen en más de un 20% de los documentos que forman un cluster, para más de 4 clusters.

Se elige inicialmente un esquema con 10 particiones considerando que se espera conseguir la detección de grandes temáticas policiales, sin interés en sub casos, por lo que 10 particiones distintas sería el máximo de casos que se esperaría obtener.

3.5. Evaluación

El desempeño de los modelos de clasificación (policial/no policial) evaluados se obtiene a partir de un esquema k-fold cross validation (con k=10) y se comparan las medidas de recall, precisión, F-measure y accuracy.

Para la evaluación de la aplicación de clustering se utiliza un criterio interno (índice Davies Bouldin) para encontrar el número óptimo de particiones para cada una de las fuentes de noticias. Se espera obtener un mismo número de particiones para las cuatro fuentes o, en el peor de los casos, un número similar de clusters. A partir de las palabras que caracterizan cada cluster se verifican que los temas tratados en los distintos medios sean similares y fácilmente identificables. Finalmente se verifica la calidad predictiva de un modelo para la clasificación de noticias dentro de temáticas policiales a partir de los resultados obtenidos del clustering, utilizando las medidas de desempeño clásicas para aprendizaje supervisado.

3.6. Desarrollo

Luego de identificadas las distintas temáticas dentro de las noticias policiales se crean diferentes mecanismos para explorar los resultados obtenidos a través de herramientas y datos externos que permiten apoyar el análisis.

Las noticias de cada una de las temáticas policiales son consolidadas en bases de datos independientes, a partir de las cuales se desarrollan herramientas útiles para su caracterización y que, a su vez, sirven como método de validación de consistencia del conocimiento detectado. Para la caracterización de una temática específica se utiliza como preprocesamiento solo la lista de stop-word y el stemming (no se utiliza ningún tipo de selección de atributos), con el fin de dar una descripción más real de la agrupación encontrada y verificar que la selección de atributos utilizada durante el algoritmo de clustering no haya orientado erróneamente la identificación de temáticas policiales. Cada temática es caracterizada a través de una herramienta de visualización conocida como tag cloud, descrita en la sección 2.3.9.1., y la identificación de reglas de asociación sencillas, descrita en la sección 2.3.9.2.

Se estudia la distribución de las distintas temáticas policiales dentro de cada región durante el período de estudio. Luego se utilizan las estadísticas de los casos policiales reales[5] asociadas a cada una de las temáticas policiales identificadas, primero se realiza una comparación del número casos y del número de noticias según la cantidad de habitantes por región para luego analizar una posible relación lineal entre el número de casos y el número de noticias mensuales.

Finalmente se construye un prototipo de una herramienta de visualización de datos georreferenciados utilizando como base el API de JavaScript de Google Maps, el que

[5] Se obtienen las estadísticas de los casos reales de delitos a través de la solicitud de información pública a Carabineros de Chile (Ley 20285 de Transparencia) y a través de información disponible en el sitio web de la Subsecretaría de Prevencion del Delito del Gobierno de Chile:
http://www.seguridadpublica.gov.cl/delitos_de_mayor_connotacion_social.html

permite insertar un mapa de Google en páginas web. Las principales tareas para realizar el prototipo son:

- Crear una página web que aloje al mapa y capture los datos georreferenciados (utilizando html y php).
- Delimitar cada región dentro del mapa.
- Capturar los datos georreferenciados desde un archivo Excel.
- Definir un criterio para clasificar los datos, asignando un color característico para cada clase.

4. Resultados y análisis

4.1. Identificación de noticias policiales

A partir de la base de datos formada por las noticias de la sección nacional de la fuente Beta, se estudian modelos para la clasificación de noticias entre Policiales y No Policiales. En el Cuadro 5 se muestran las palabras que aparecen en mayor cantidad de noticias para cada clase.

Palabras	Word Frequency	
	Clase Policial	Clase No Policial
carabinero	540	43
detener	452	17
robar	299	3
pdi	288	5
comuna	284	96
incendiar	197	5
morir	187	46
persona	181	81
sujeto	161	1
vivienda	156	125
brigada	140	1
vehículo	132	26
delincuente	130	2
herir	128	7
investigar	127	65
menor	127	43
asaltar	127	0
personal	115	19
calle	110	28
sector	109	107

Palabras	Word Frequency	
	Clase Policial	Clase No Policial
diputado	11	567
gobierno	26	408
presidente	18	375
ministro	17	346
proyecto	7	308
educar	20	290
senador	2	283
udi	1	253
nuevo	48	231
presupuestar	2	229
ley	19	213
realizar	91	195
rechazar	5	188
partido	7	187
cámara	2	185
aprobar	6	179
comisión	2	166
presentar	24	156
ppd	2	155
votar	9	152

Cuadro 5: Palabras más frecuentes para la clase Policial y la clase No Policial
Fuente: Elaboración propia

Antes de la aplicación de modelos de clasificación se realiza una selección de atributos usando como criterios que las palabras estén presentes en más de un 0,5% de los documentos y que simultáneamente estén entre los valores más altos para Chi-Cuadrado e Information Gain (Cuadro 6).

Chi-Cuadrado		
carabinero	incendiar	herir
detener	senador	cámara
pdi	ministro	udi
diputado	delincuente	calle
robar	proyecto	homicidio
comuna	madrugada	identificar
brigada	personal	imputar
sujeto	asaltar	partido
presidente	morir	delito
gobierno	arrestar	banda

Information Gain		
carabinero	gobierno	partido
detener	ministro	personal
diputado	proyecto	herir
pdi	incendiar	morir
robar	delincuente	presupuestar
brigada	asaltar	homicidio
sujeto	cámara	comisión
presidente	udi	banda
senador	arrestar	educar
comuna	madrugada	identificar

Cuadro 6: Palabras con los más altos valores de Chi-Cuadrado e Information Gain.
Fuente: Elaboración propia

Se establecen dos mecanismos para estudiar los modelos de clasificación, utilizando primero las 400 y luego las 800 palabras con mayores valores para Chi-cuadrado e Information Gain. Los modelos estudiados son Naive Bayes y K-nn (estudiados para distintos valores de K).

Modelo	Recall	Precisión	F-measure	Accuracy
Naive Bayes (400 Palabras)	93,79% +/- 1,80%	79,94% +/- 1,89%	86,29% +/- 1,33%	88,18% +/- 1,51%
Naive Bayes (800 Palabras)	91,03% +/- 2,51%	81,80% +/- 1,95%	86,13% +/- 1,23%	88,37% +/- 1,39%
K-nn (400 Palabras y K=11)	91,73% +/- 2,14%	92,54% +/- 2,66%	92,10% +/- 1,57%	93,73% +/- 1,40%
K-nn (800 Palabras y K=19)	92,28% +/- 2,07%	91,37% +/- 2,02%	91,78% +/- 0,92%	93,43% +/- 0,97%

Cuadro 7: Medidas de desempeño para distintos modelos de clasificación
Fuente: Elaboración propia

Dentro de los modelos y esquemas evaluados se selecciona el modelo K-nn con K=11 utilizando 400 palabras, por tener los valores más altos para precisión, accuracy y F-measure. El desempeño de los modelos K-nn para distintos valores de k se muestra en la Figura 3 y Figura 4.

El modelo de clasificación seleccionado es aplicado a las bases de datos formadas por las noticias de las secciones nacionales de las fuentes Gamma y Delta. Para la fuente Gamma se clasifican 2.440 noticias como policiales (de un total de 6.286), mientras que para la fuente Delta 2.767 (de un total de 5.937). De esta forma se logra formar una base de datos únicamente con noticias policiales, con un total de 14.045 noticias.

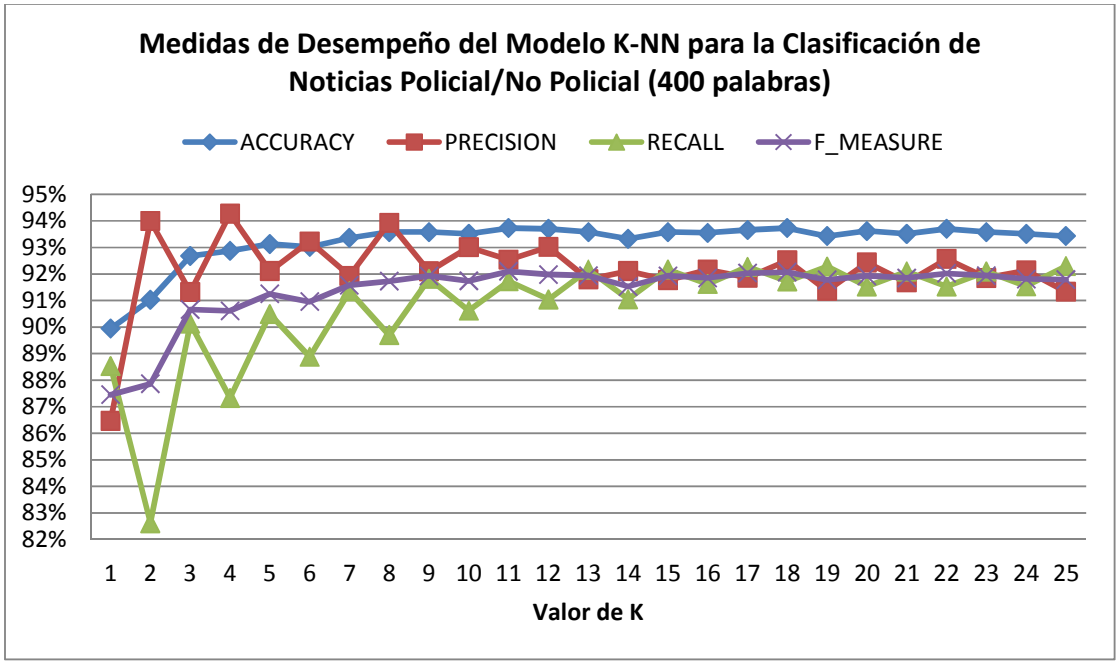

Figura 3: Medidas de desempeño modelo K-nn (con 400 palabras)
Fuente: Elaboración propia

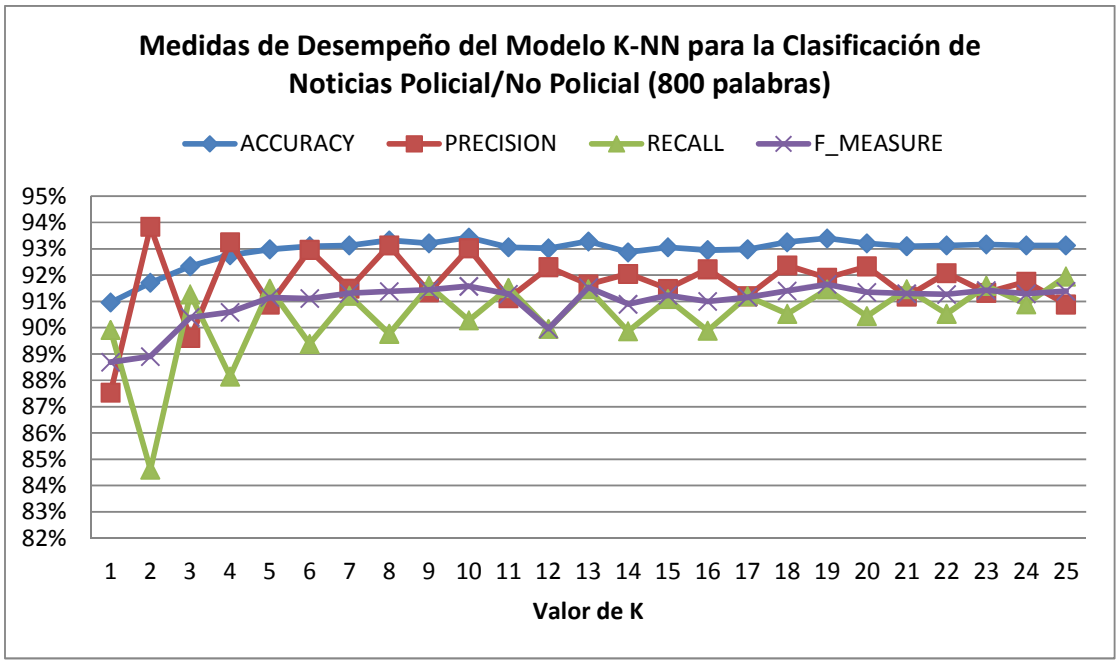

Figura 4: Medidas de desempeño modelo K-nn (con 800 palabras)
Fuente: Elaboración propia

4.2. Identificación de temáticas policiales

A partir de las bases de noticias policiales de cada uno de los cuatro medios de prensa se realiza el procedimiento para identificar las temáticas policiales tratadas en cada medio. El primer paso es realizar una selección de atributos usando como criterios que las palabras estén presentes en más de un 0,1% de los documentos y que simultáneamente estén entre los valores más altos para TVQ, TV y TC. Para cada fuente de noticias se seleccionan las 1.000 palabras con los más altos valores según cada uno de los tres criterios. En el Cuadro 8 y Cuadro 9 se muestran las palabras con los mayores valores según cada criterio.

Se aplica el algoritmo de clustering K-means, con un valor de k igual a diez, a cada una de las fuentes de noticias policiales. Antes de identificar las temáticas, se estudian aquellas palabras muy frecuentes y que no entregan valor para distinguir una temática de otra. El criterio a utilizar es: si una palabra aparece en más de una 10% de los documentos que forman un cluster y esto se repite para más de 7 clusters o bien si una palabra aparece en más de una 20% de los documentos de un cluster y esto se repite para más de 4 clusters, entonces la palabra se elimina del análisis. De esta forma para cada fuente de noticias policiales se detectan las palabras muy frecuentes pero sin poder de discriminación, de las cuales 28 se identifican en las cuatro fuentes (Cuadro 10).

Fuente Alfa		
TVQ	**TV**	**TC**
detener	detener	detener
carabinero	carabinero	robar
robar	robar	sujeto
encontrar	encontrar	accidente
menor	menor	menor
mujer	mujer	carabinero
vehículo	joven	joven
lugar	sujeto	mujer
persona	vehículo	vehículo
sujeto	accidente	encontrar
joven	lugar	lesionar
sector	sector	sector
accidente	persona	hombre
investigar	pdi	persona
policial	investigar	lugar

Fuente Beta		
TVQ	**TV**	**TC**
carabinero	detener	robar
detener	carabinero	detener
robar	robar	pdi
vehículo	pdi	sujeto
pdi	vehículo	vehículo
persona	persona	carabinero
accidente	sujeto	menor
menor	droga	herir
droga	menor	droga
jefe	accidente	accidente
sujeto	mujer	persona
delito	delito	morir
mujer	herir	mujer
tránsito	encontrar	imputar
herir	morir	encontrar

Cuadro 8: Palabras con mayores valores de TVQ, TV y TC de las fuentes Alfa y Beta
Fuente: Elaboración propia

Fuente Gamma		
TVQ	**TV**	**TC**
carabinero	carabinero	accidente
marchar	vehículo	robar
vehículo	accidente	vehículo
accidente	marchar	detener
robar	robar	incendiar
persona	detener	vivienda
detener	persona	carabinero
vivienda	vivienda	sujeto
calle	incendiar	morir
sector	calle	menor
estudiante	estudiante	persona
hospital	encontrar	policial
herir	menor	calle
colisionar	sector	joven
caso	policial	herir

Fuente Delta		
TVQ	**TV**	**TC**
carabinero	carabinero	detener
vivienda	detener	robar
detener	marchar	vehículo
marchar	robar	carabinero
vehículo	vivienda	menor
robar	vehículo	vivienda
calle	menor	marchar
menor	encontrar	incendiar
encontrar	persona	accidente
persona	calle	sujeto
estudiante	accidente	estudiante
accidente	estudiante	persona
incendiar	incendiar	encontrar
investigar	morir	morir
delito	investigar	herir

Cuadro 9: Palabras con mayores valores de TVQ, TV y TC de las fuentes Gamma y Delta
Fuente: Elaboración propia

Palabras Frecuentes			
carabinero	persona	investigar	ubicar
comuna	realizar	lograr	antecedente
encontrar	registrar	menor	señalar
informar	sector	ocurrir	vehículo
llegar	vivienda	personal	víctima
lugar	calle	policial	identificar
momento	interior	detener	sujeto

Cuadro 10: Palabras frecuentes pero poco útiles para discriminar particiones de noticias
Fuente: Elaboración propia

Luego, por inspección, se identifica una partición caracterizada por palabras de índole judicial, en las cuatro fuentes de noticias, la cual no aportaría valor para el estudio particular, el cual se enfoca en los tipos de delitos (o bien, en los motivos que generaron alguna acción judicial). El listado de palabras que caracterizan una partición judicial (Cuadro 11) es eliminado del análisis de clustering. Así mismo se detectan otras particiones cuyas palabras que las caracterizan no muestran sentido alguno, las que son igualmente eliminadas del análisis, pues pueden alterar los resultados (ejemplos de estos casos se muestran en el Cuadro 12).

Partición temática “judicial”			
fiscal	penal	juzgado	pedir
tribunal	garantía	abogar	establecer
imputar	juicio	hecho	decidir
caso	prisión	juez	cometer
acusar	audiencia	declarar	cárcel
fiscalía	oral	libertad	delito
público	pena	cargo	disponer
ministerio	preventivo	judicial	investigador
condenar	solicitar	justicia	detective
formalizar	presentar	denunciar	ley

Cuadro 11: Palabras que caracterizan una partición “judicial”
Fuente: Elaboración propia

Partición 1		Partición 2		Partición 3	
pasar	bueno	habitante	principal	analizar	razón
sumar	raíz	contener	asociar	comenzar	sistema
poder	dicha	grado	iniciar	involucrar	comunicar
contactar	contener	unir	minutos	ingresar	intervenir
entrevistar	citado	alcanzar	informe	después	enviar
análisis	orden	mover	durar	previo	gente
específico	informe	implicar	posterior	retirar	reparar
cambiar	dato	responsable	activar	detallar	constante

Cuadro 12: Palabras que caracterizan particiones sin contenido claro
Fuente: Elaboración propia

Ahora ya se está en condiciones de realizar K-means con las palabras definitivas que se utilizarán en el estudio de identificación de temáticas policiales, para lo cual se evalúa el mejor esquema (número de clusters) para cada una de las fuentes de noticias policiales utilizando el índice Davies-Bouldin. Para las cuatro fuentes se obtiene que el mejor esquema es la partición de las noticias en 7 clusters, dado que los menores valores de Davies-Bouldin se alcanza para K=7 (Figura 5). Por inspección de las particiones determinadas, se observa un sentido claro y fácilmente identificable de cada una de ellas a través de las palabras que las caracterizan, además de que las siete mismas temáticas pueden ser identificadas en las cuatro fuentes. Las distintas temáticas representadas por cada cluster se presentan en el Cuadro 13, en el cual cada cluster se presenta con un nombre que hace referencia a su contenido.

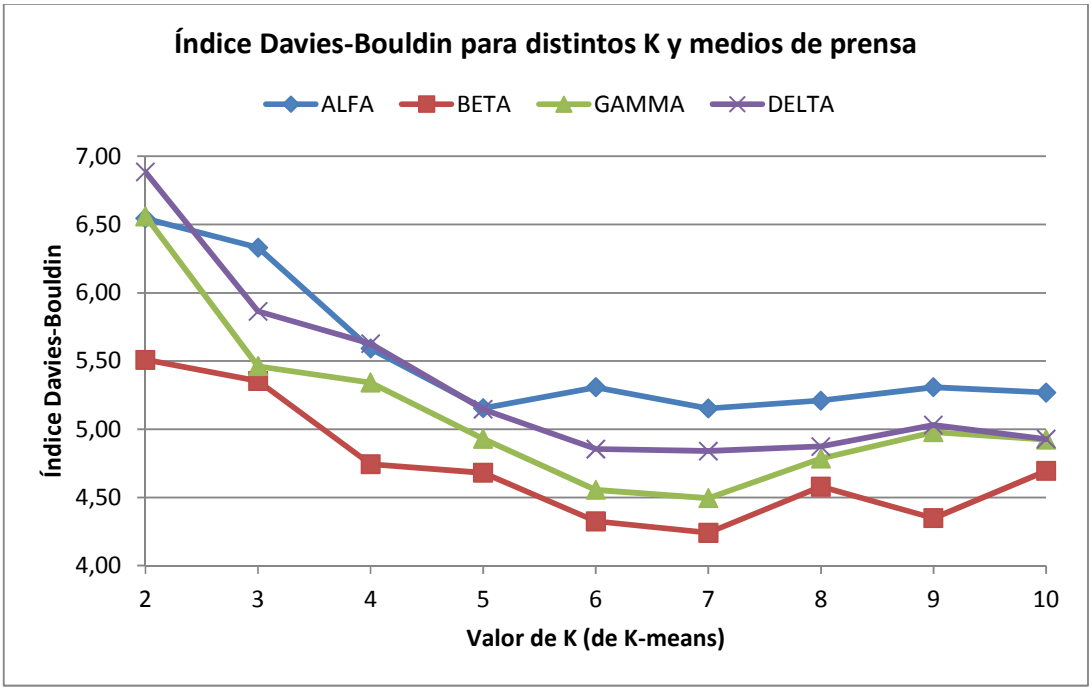

Figura 5: Índice Davies-Bouldin para distintos K y distintos medios de prensa
Fuente: Elaboración propia

Temática	N° Noticias	% Noticias
"Delitos Sexuales"	1.185	8,4%
"Incendios"	1.223	8,7%
"Drogas"	1.455	10,4%
"Disturbios"	2.196	15,6%
"Homicidios"	2.407	17,1%
"Tránsito"	2.572	18,3%
"Robos"	3.007	21,4%
Total	14.045	100%

Cuadro 13: Distribución temáticas policiales en el período de estudio
Fuente: Elaboración propia

En la Figura 6 se puede apreciar que las distintas temáticas no tienen necesariamente la misma cobertura en los distintos medios de prensa durante el período de estudio. Por otro lado, en el Figura 7 se puede observar que ciertas temáticas policiales tuvieron mayor cobertura durante algunos meses, principalmente esta situación destaca para la temática disturbios y su cobertura del mes de agosto que coincide con el acontecer nacional de movilizaciones estudiantiles (cuyo auge de manifestaciones se vivió en agosto). Otros cambios en la cobertura pueden estar relacionados con situaciones climáticas como en el caso de la temática incendios cuya mayor cobertura fue en diciembre (comienzo del verano y aumento de temperaturas) o

bien con festividades como en el caso de la temática tránsito que presenta una mayor cobertura durante septiembre que se caracteriza por la celebración de fiestas patrias y el traslado de gran cantidad de personas.

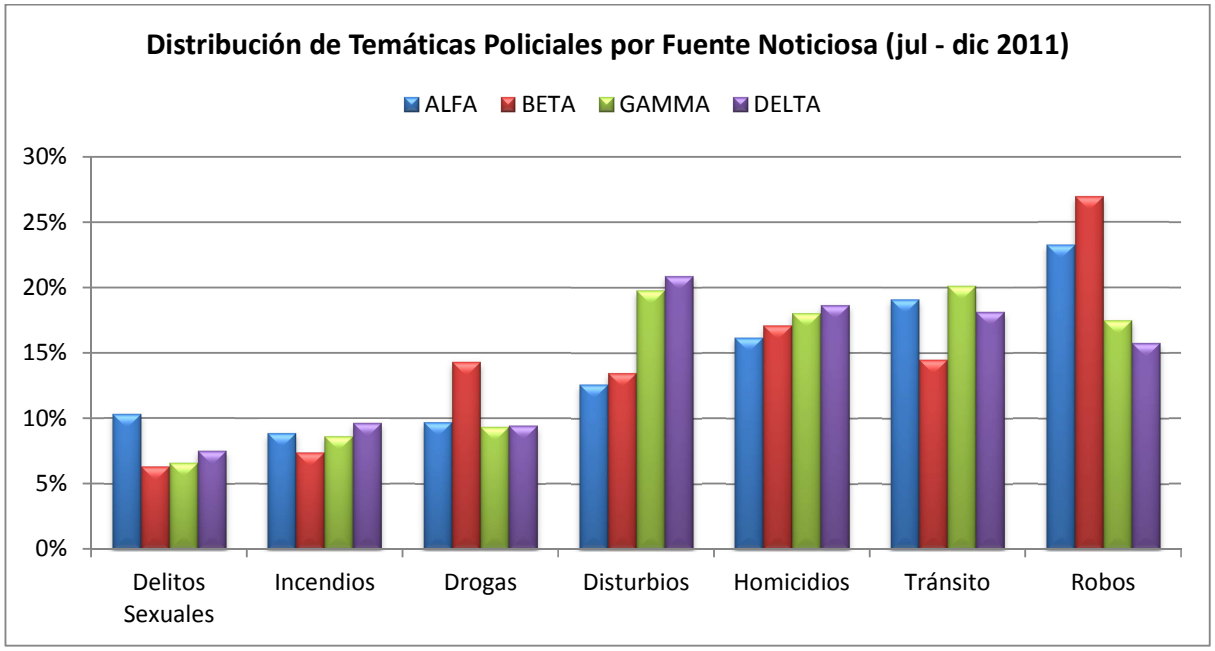

Figura 6: Distribución de las temáticas según fuente noticiosa
Fuente: Elaboración propia

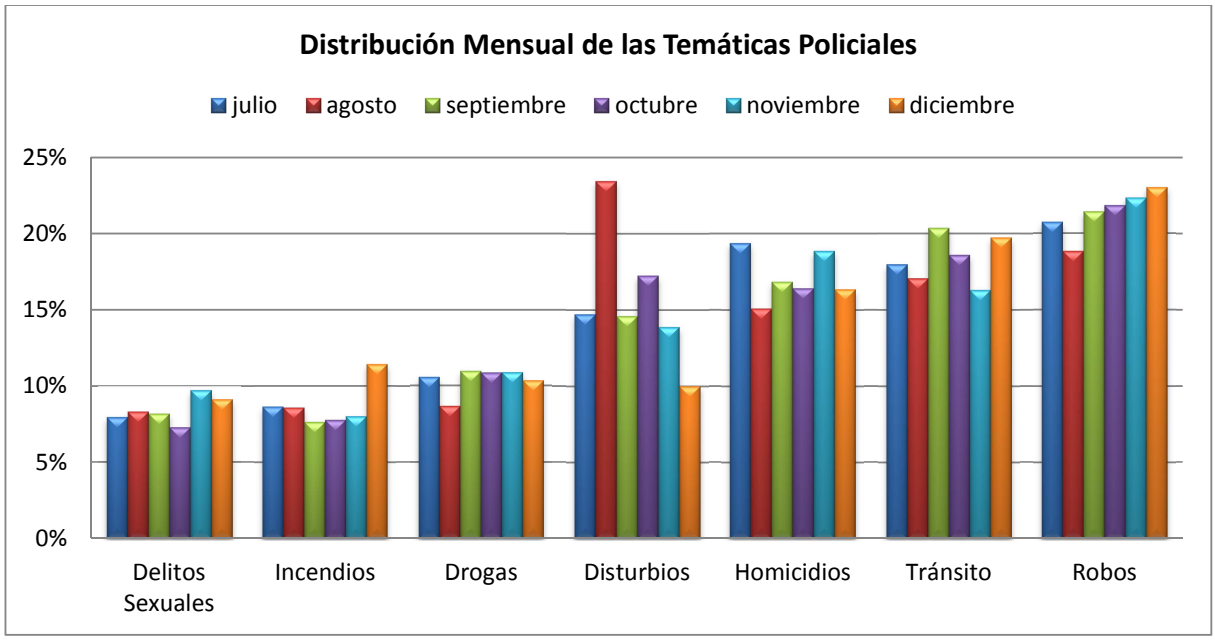

Figura 7: Distribución mensual de las temáticas policiales
Fuente: Elaboración propia

4.2.1. Caracterización de la temática drogas

En el Cuadro 14 se muestran las palabras con mayor relevancia de la temática drogas para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se logra observar que de las 20 palabras más representativas del cluster, el 75% se repiten en las cuatro fuentes.

Fuente Alfa	Fuente Beta	Fuente Gamma	Fuente Delta
droga	droga	droga	incautar
incautar	incautar	incautar	droga
marihuana	cocaína	narcotráfico	cocaína
cocaína	kilo	cocaína	peso
operativo	peso	kilo	marihuana
base	gramo	marihuana	kilo
gramo	marihuana	gramo	tráfico
tráfico	infraccionar	peso	base
kilo	base	base	gramo
comercializar	decomisar	decomisar	operativo
decomisar	dosis	operativo	comercializar
pasta	pasta	pasta	millón
peso	tráfico	ocultar	dosis
antinarcótico	arma	comercializar	arma
portar	sustancia	millón	pasta
dosis	ocultar	dosis	portar
millón	millón	tráfico	banda
dedicar	antinarcótico	infraccionar	dedicar
infraccionar	comercializar	paquete	antinarcótico
vender	operativo	antinarcótico	dinero

Cuadro 14: Palabras más relevantes del cluster drogas
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre drogas de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática drogas no pierde su caracterización sin la selección de atributos y resulta fácilmente identificable a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 15), un número pequeño de reglas de asociación (Cuadro 16) y un tag cloud (Figura 8).

Palabras	Document Frequency	% de noticias
detener	1133	78%
droga	1040	71%
incautar	823	57%
carabinero	659	45%
pdi	621	43%
cocaína	590	41%
realizar	575	40%
investigar	566	39%
policial	563	39%
encontrar	553	38%

Palabras	Document Frequency	% de noticias
control	539	37%
personal	531	36%
sujeto	531	36%
disponer	526	36%
marihuana	514	35%
policía	486	33%
vivienda	464	32%
efectivo	462	32%
kilo	461	32%
brigada	457	31%

Cuadro 15: Palabras con mayor document frequency del cluster drogas
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
prensado	marihuana	8%	100%
pasta	base	23%	100%
kilo, avaluar	millón	6%	98%
cannabis	sativa	8%	97%
clorhidrato	cocaína	13%	97%
detener, fuego	arma	7%	96%
droga, antinarcótico	brigada	16%	95%
detener, control, ley	infraccionar	11%	91%
paquete	kilo	11%	89%
detener, encontrar, allanar	vivienda	5%	88%

Cuadro 16: Reglas de asociación detectadas dentro de la temática drogas
Fuente: Elaboración propia

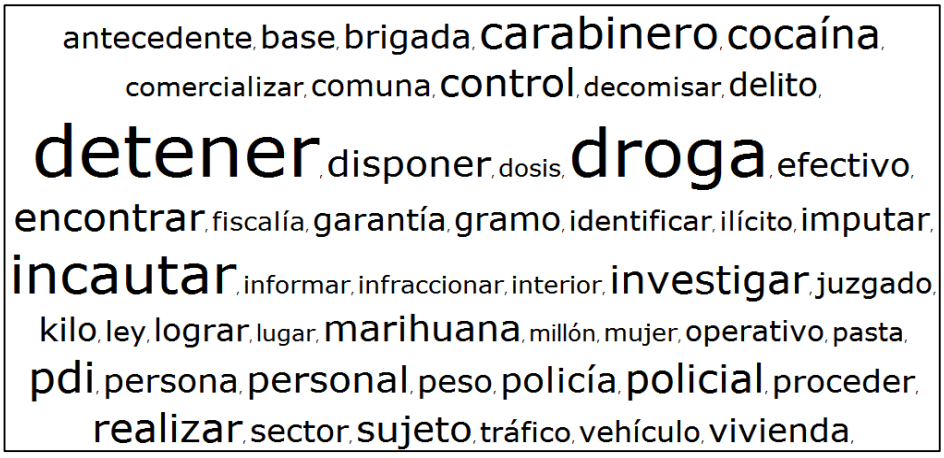

Figura 8: Tag cloud para la temática drogas
Fuente: Elaboración propia

4.2.2. Caracterización de la temática robos

En el Cuadro 17 se muestran las palabras con mayor relevancia de la temática robos para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se observa que de las 20 palabras más representativas del cluster, el 55% se repiten en las cuatro fuentes.

Fuente ALFA	Fuente BETA	Fuente GAMMA	Fuente DELTA
robar	robar	robar	robar
delincuente	delincuente	delincuente	delincuente
especie	peso	millón	millón
peso	millón	asaltar	asaltar
dinero	especie	antisocial	antisocial
millón	asaltar	huir	especie
asaltar	arma	peso	arma
antisocial	huir	especie	peso
arma	dinero	arma	intimidar
huir	fuego	dinero	dinero
desconocido	intimidar	intimidar	fuego
recuperar	banda	banda	grupo
avaluar	apropiar	fuego	huir
autor	recuperar	grupo	banda
ladrón	cajero	cajero	cajero
fugar	escapar	encargar	paradero
dueño	antisocial	atracar	encargar
intimidar	maleante	escapar	escapar
paradero	incautar	perpetrar	capturar
grupo	automático	buscar	cometer

Cuadro 17: Palabras más relevantes del cluster robos
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre robos de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática robos no pierde su caracterización sin la selección de atributos y resulta fácilmente identificable a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 18), un número pequeño de reglas de asociación (Cuadro 19) y un tag cloud (Figura 9).

Palabras	Document Frequency	% de noticias	Palabras	Document Frequency	% de noticias
robar	2522	84%	vehículo	1061	35%
carabinero	1938	64%	comuna	1010	34%
detener	1637	54%	encontrar	1005	33%
sujeto	1618	54%	peso	940	31%
delincuente	1387	46%	antecedente	939	31%
calle	1246	41%	llegar	929	31%
lugar	1187	39%	personal	929	31%
ubicar	1145	38%	millón	899	30%
policial	1142	38%	lograr	898	30%
delito	1103	37%	policía	890	30%

Cuadro 18: Palabras con mayor document frequency del cluster robos
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
dispensador	cajero, automático	7%	97%
delincuente, automático	robar, cajero	7%	96%
fuego	arma	15%	96%
lugar, delito	robar	12%	95%
comisaría	carabinero	14%	94%
intimidar, fuego	robar, arma	7%	90%
detener, audiencia	control	6%	86%
carabinero, delincuente, vehículo	sujeto	6%	81%
violento	asaltar	5%	80%
robar, peso, especie	millón	8%	80%

Cuadro 19: Reglas de asociación detectadas dentro de la temática robos
Fuente: Elaboración propia

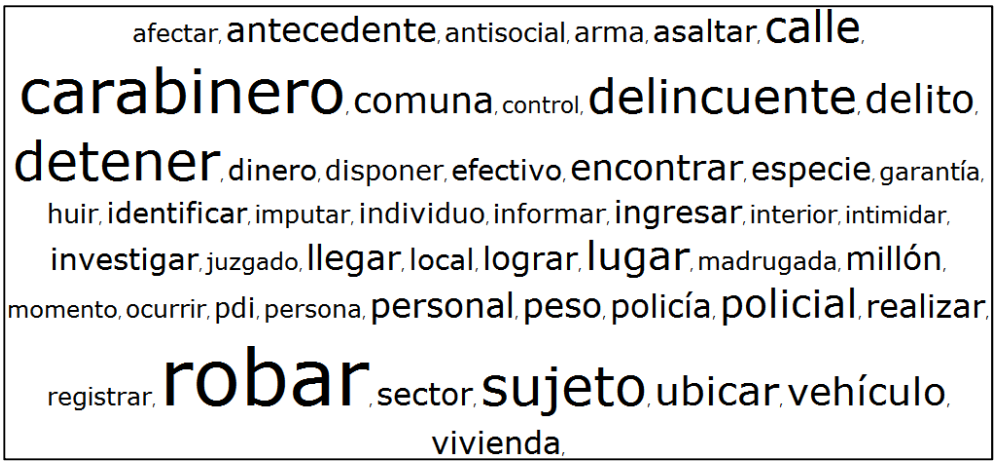

Figura 9: Tag cloud para la temática robos
Fuente: Elaboración propia

4.2.3. Caracterización de la temática delitos sexuales

En el Cuadro 20 se muestran las palabras con mayor relevancia de la temática delitos sexuales para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se observa que de las 20 palabras más representativas del cluster, el 40% se repiten en las cuatro fuentes.

Fuente ALFA	Fuente BETA	Fuente GAMMA	Fuente DELTA
sexual	sexual	sexual	abusar
niño	abusar	abusar	sexual
hijo	niño	niño	niño
abusar	violar	violar	violar
agredir	hijo	aprovechar	pareja
madre	madre	pequeño	cometer
violar	brisexme	pareja	agredir
padre	aprovechar	reiterar	inicial
pareja	pareja	agredir	amenazar
agresor	padre	atacar	infantil
pequeño	pequeño	relación	pequeño
golpear	amenazar	indagatorio	aprovechar
violencia	reiterar	amenazar	autor
tomar	relación	infantil	manifestar
aprovechar	infantil	golpear	violencia
internar	agredir	ataque	ataque
buscar	colegio	tomar	frustrar
atacar	descubierto	concurrir	homicidio
reiterar	reconocer	acreditar	maltratar
intrafamiliar	buscar	iglesia	tomar

Cuadro 20: Palabras más relevantes del cluster delitos sexuales
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre delitos sexuales de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática delitos sexuales no pierde su caracterización sin la selección de atributos y resulta posible identificarla a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 21), un número pequeño de reglas de asociación (Cuadro 22) y un tag cloud (Figura 10).

Palabras	Document Frequency	% de noticias
menor	623	53%
detener	595	50%
delito	530	45%
sexual	526	44%
vivienda	497	42%
denunciar	462	39%
sujeto	460	39%
abusar	455	38%
carabinero	446	38%
encontrar	438	37%

Palabras	Document Frequency	% de noticias
víctima	435	37%
investigar	431	36%
niño	418	35%
realizar	379	32%
pdi	376	32%
hombre	375	32%
acusar	349	29%
hijo	344	29%
identificar	339	29%
mujer	338	29%

Cuadro 21: Palabras con mayor document frequency del cluster delitos sexuales
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
imputar, prisión	preventivo	6%	97%
intrafamiliar	violencia	7%	96%
sexual, brigada	delito	17%	94%
brisexme	delito, sexual, brigada	8%	92%
sexual, niño, acusar	abusar	9%	91%
abusar	sexual	34%	89%
delito, sexual, sujeto	menor	16%	85%
sexual, acusar, inicial	identificar	5%	83%
tocar	sexual, abusar	8%	81%
detener, sexual, brigada	sujeto	12%	80%

Cuadro 22: Reglas de asociación detectadas dentro de la temática delitos sexuales
Fuente: Elaboración propia

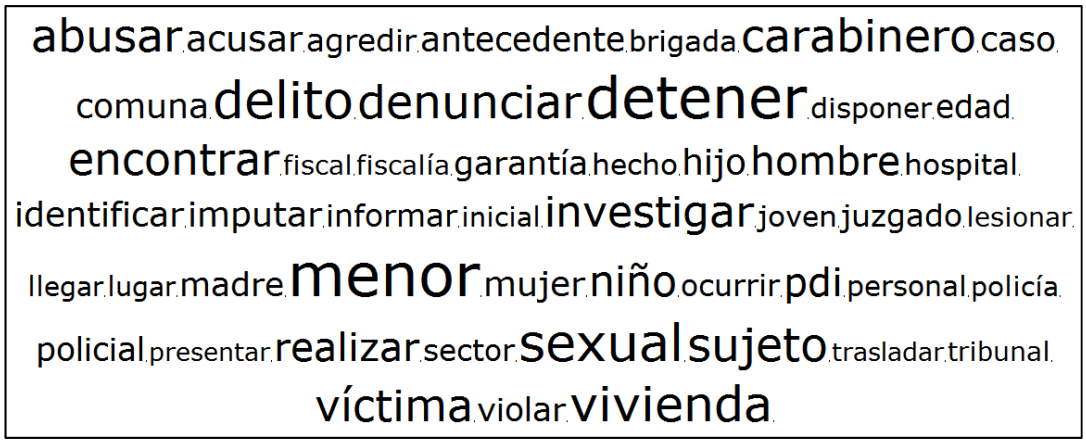

Figura 10: Tag cloud para la temática delitos sexuales
Fuente: Elaboración propia

4.2.4. Caracterización de la temática homicidios

En el Cuadro 23 se muestran las palabras con mayor relevancia de la temática homicidios para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se observa que de las 20 palabras más representativas del cluster, el 40% se repiten en las cuatro fuentes.

Fuente ALFA	Fuente BETA	Fuente GAMMA	Fuente DELTA
morir	homicidio	homicidio	homicidio
homicidio	morir	disparar	fallecer
cuerpo	fallecer	arma	disparar
fallecer	arma	fallecer	arma
autor	disparar	autor	autor
crimen	cuerpo	cuerpo	cuerpo
arma	crimen	crimen	bala
hallazgo	autor	asesinar	crimen
cadáver	vida	pareja	impactar
asesinar	hallar	bala	testigo
disparar	hijo	ataque	balear
hijo	cadáver	hallar	incidente
buscar	cabeza	cabeza	tomar
trabajador	bala	testigo	fuego
tomar	impactar	agredir	periciar
periciar	agredir	buscar	asesinar
noche	buscar	cadáver	buscar
hallar	pareja	matar	pareja
deceso	periciar	grupo	tórax
suicidar	grupo	tomar	grupo

Cuadro 23: Palabras más relevantes del cluster homicidios
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre homicidios de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática homicidios no pierde su caracterización sin la selección de atributos y resulta posible identificarla a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 24), un número pequeño de reglas de asociación (Cuadro 25) y un tag cloud (Figura 11).

Palabras	Document Frequency	% de noticias
morir	1344	56%
encontrar	1279	53%
carabinero	1134	47%
homicidio	1107	46%
investigar	992	41%
pdi	939	39%
víctima	932	39%
lugar	891	37%
fallecer	883	37%
comuna	861	36%

Palabras	Document Frequency	% de noticias
identificar	861	36%
vivienda	818	34%
realizar	816	34%
ocurrir	788	33%
brigada	785	33%
hombre	785	33%
persona	769	32%
detener	766	32%
policial	766	32%
cuerpo	765	32%

Cuadro 24: Palabras con mayor document frequency del cluster homicidios
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
servicio, médico	legal	14%	95%
disparar, fuego	arma	6%	95%
criminalístico	laboratorio	5%	95%
brigada	homicidio	31%	94%
quitar	vida	6%	92%
homicidio, pdi	brigada	26%	92%
turno	fiscal	5%	91%
urgencia	hospital	6%	91%
riesgo	vital	6%	90%
hallazgo	encontrar	9%	85%

Cuadro 25: Reglas de asociación detectadas dentro de la temática homicidios
Fuente: Elaboración propia

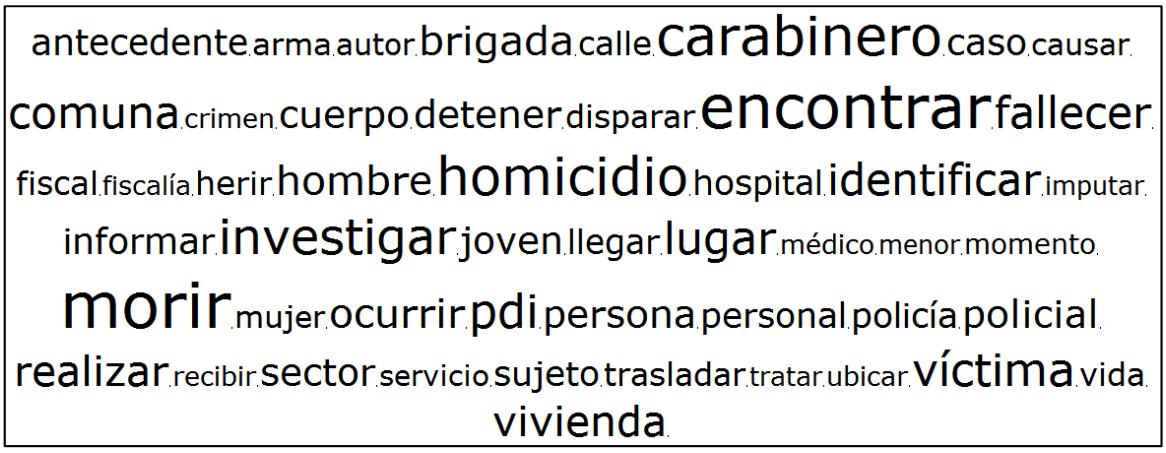

Figura 11: Tag cloud para la temática homicidios
Fuente: Elaboración propia

4.2.5. Caracterización de la temática tránsito

En el Cuadro 26 se muestran las palabras con mayor relevancia de la temática tránsito para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se observa que de las 20 palabras más representativas del cluster, el 55% se repiten en las cuatro fuentes.

Fuente ALFA	Fuente BETA	Fuente GAMMA	Fuente DELTA
accidente	accidente	accidente	accidente
conductor	tránsito	tránsito	tránsito
colisionar	colisionar	conductor	fallecer
impactar	morir	fallecer	ruta
ruta	ruta	ruta	conductor
tránsito	fallecer	colisionar	colisionar
conducir	conductor	impactar	kilómetro
bombero	kilómetro	viajar	impactar
volcar	impactar	kilómetro	viajar
fallecer	chofer	camión	camión
camión	camión	bombero	transportar
chofer	perder	bus	bus
kilómetro	siat	fatal	fatal
morir	bus	vía	conducir
camioneta	fatal	transportar	pasajero
samu	bombero	cuerpo	bombero
atropellar	carretero	siat	carretero
móvil	atropellar	manejar	vía
viajar	volcar	pasajero	volcar
máquina	viajar	atropellar	operativo

Cuadro 26: Palabras más relevantes del cluster tránsito
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre tránsito de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática tránsito no pierde su caracterización sin la selección de atributos y resulta fácilmente identificable a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 27), un número pequeño de reglas de asociación (Cuadro 28) y un tag cloud (Figura 12).

Palabras	Document Frequency	% de noticias
accidente	2031	79%
vehículo	1567	61%
carabinero	1558	61%
persona	1158	45%
lugar	1147	45%
conductor	1107	43%
lesionar	1042	41%
tránsito	1010	39%
colisionar	1000	39%
trasladar	991	39%

Palabras	Document Frequency	% de noticias
hospital	959	37%
ocurrir	927	36%
ruta	907	35%
sector	904	35%
personal	881	34%
informar	860	33%
encontrar	847	33%
resultar	821	32%
herir	818	32%
calle	790	31%

Cuadro 27: Palabras con mayor document frequency del cluster tránsito
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
hospital, riesgo	vital	6%	98%
perder, volcar	control	6%	97%
tránsito, colisionar	accidente	18%	96%
tránsito, siat	accidente, carabinero	8%	91%
impactar, control	perder	6%	90%
accidente, lugar, colisionar, impactar	vehículo	8%	90%
vehículo, leve	lesionar	6%	89%
lesionar, rescatar	bombero	5%	88%
identificar, fatal	víctima	5%	85%
trasladar, herir	hospital	15%	81%

Cuadro 28: Reglas de asociación detectadas dentro de la temática tránsito
Fuente: Elaboración propia

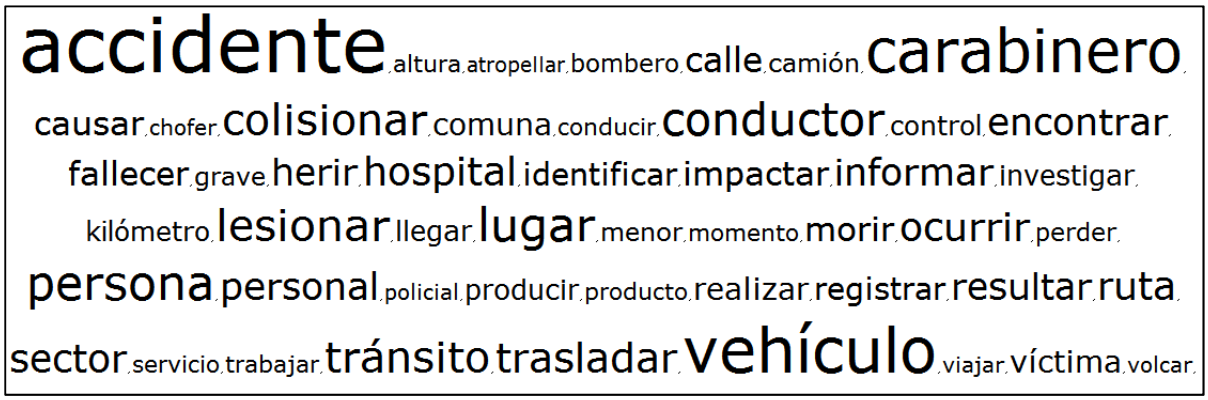

Figura 12: Tag cloud para la temática tránsito
Fuente: Elaboración propia

4.2.6. Caracterización de la temática disturbios

En el Cuadro 29 se muestran las palabras con mayor relevancia de la temática disturbios para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se observa que de las 20 palabras más representativas del cluster, el 65% se repiten en las cuatro fuentes.

Fuente ALFA	Fuente BETA	Fuente GAMMA	Fuente DELTA
grupo	grupo	educar	estudiante
estudiante	especial	estudiante	grupo
especial	fuerza	grupo	marchar
tomar	estudiante	manifestar	manifestar
incidente	lanzar	marchar	manifestante
universidad	universidad	especial	tomar
fuerza	incidente	tomar	fuerza
manifestar	tránsito	movilizar	movilizar
lanzar	encapuchado	universidad	estudiantil
barricada	manifestante	incidente	especial
encapuchado	barricada	estudiantil	incidente
enfrentar	marchar	fuerza	tránsito
dañar	carro	manifestante	universidad
manifestante	manifestar	tránsito	plaza
tránsito	tomar	protestar	enfrentar
marchar	enfrentar	encapuchado	educar
ataque	plaza	lanzar	alumno
alumno	educar	enfrentar	convocar
estudiar	estudiantil	alumno	autorizar
establecimiento	movilizar	barricada	barricada

Cuadro 29: Palabras más relevantes del cluster disturbios
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre disturbios de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática disturbios no pierde su caracterización sin la selección de atributos y resulta posible identificarla a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 30), un número pequeño de reglas de asociación (Cuadro 31) y un tag cloud (Figura 13).

Palabras	Document Frequency	% de noticias
carabinero	1610	73%
grupo	906	41%
detener	895	41%
estudiante	860	39%
calle	833	38%
persona	772	35%
policial	722	33%
realizar	717	33%
lugar	700	32%
encontrar	679	31%

Palabras	Document Frequency	% de noticias
sector	655	30%
informar	649	30%
llegar	644	29%
especial	638	29%
personal	587	27%
manifestar	585	27%
marchar	578	26%
tomar	567	26%
fuerza	559	25%
joven	548	25%

Cuadro 30: Palabras con mayor document frequency del cluster disturbios
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
lanzaagua	carro	6%	100%
autorizar, intendencia	marchar	6%	99%
molotov	bomba	8%	97%
especial, fuerza	carabinero	20%	96%
carabinero, secundario	estudiante	7%	90%
calle, fuerza	especial	12%	88%
carabinero, persona, desorden	detener	5%	86%
carabinero, interrumpir	tránsito	6%	85%
estudiante, desalojar	tomar	6%	83%
carabinero, tránsito, encapuchado	barricada	5%	80%

Cuadro 31: Reglas de asociación detectadas dentro de la temática disturbios
Fuente: Elaboración propia

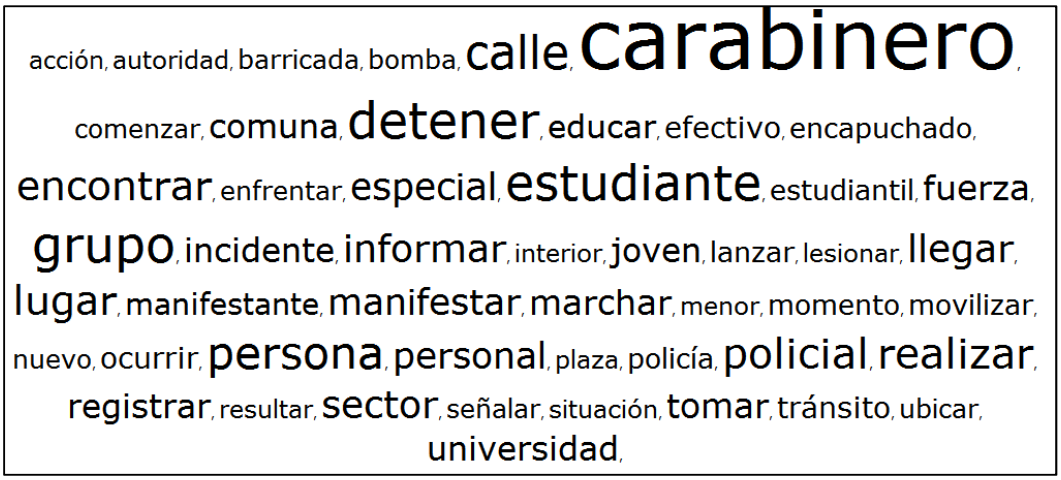

Figura 13: Tag cloud para la temática disturbios
Fuente: Elaboración propia

4.2.7. Caracterización de la temática incendios

En el Cuadro 32 se muestran las palabras con mayor relevancia de la temática incendios para cada una de las fuentes de noticias, determinadas a través del vector de centroides encontrado por el algoritmo de clustering. Se observa que de las 20 palabras más representativas del cluster, el 70% se repiten en las cuatro fuentes.

Fuente ALFA	Fuente BETA	Fuente GAMMA	Fuente DELTA
incendiar	incendiar	incendiar	incendiar
bombero	bombero	bombero	bombero
siniestro	siniestro	siniestro	siniestro
llama	compañía	fuego	fuego
fuego	llama	llama	controlar
compañía	fuego	compañía	llama
destruir	controlar	voluntario	compañía
inmueble	voluntario	controlar	consumir
controlar	inmueble	cuerpo	voluntario
voluntario	destruir	consumir	cuerpo
dañar	morir	destruir	forestal
consumir	damnificado	combatir	comandante
concurrir	fallecer	propagar	combatir
cuerpo	cuerpo	fallecer	hectárea
propagar	extinguir	rápido	completar
material	propagar	concurrir	fallecer
damnificado	material	inmueble	material
piso	concurrir	origen	destruir
rápido	originar	material	inmueble
quemar	consumir	evacuar	propagar

Cuadro 32: Palabras más relevantes del cluster incendios
Fuente: Elaboración propia

Una vez consolidadas todas las noticias sobre incendios de los distintos medios de prensa se realiza una caracterización de la temática utilizando solo la lista de stop-word y stemming (no se utiliza ningún tipo de selección de atributos).

Por inspección se considera que la temática incendios no pierde su caracterización sin la selección de atributos y resulta fácilmente identificable a través del uso de un listado de palabras que más se repiten entre las noticias (Cuadro 33), un número pequeño de reglas de asociación (Cuadro 34) y un tag cloud (Figura 14).

Palabras	Document Frequency	% de noticias
incendiar	1024	84%
bombero	930	76%
lugar	670	55%
siniestro	663	54%
afectar	594	49%
vivienda	586	48%
fuego	572	47%
llama	560	46%
compañía	506	41%
sector	506	41%

Palabras	Document Frequency	% de noticias
persona	497	41%
calle	481	39%
ubicar	462	38%
informar	437	36%
causar	429	35%
encontrar	416	34%
trabajar	412	34%
controlar	411	34%
comuna	405	33%
llegar	392	32%

Cuadro 33: Palabras con mayor document frequency del cluster incendios
Fuente: Elaboración propia

Antecedente	Consecuente	Soporte	Confianza
siniestro	incendiar	53%	98%
compañía	bombero	40%	97%
fatal	víctima	6%	97%
ligero	incendiar, vivienda, material	10%	94%
incendiar, llama, llegar, propagar	lugar	5%	90%
damnificado	vivienda	13%	90%
bombero, vivienda, propagar	llama	10%	89%
conaf	incendiar, forestal	5%	86%
bombero, afectar, destruir	siniestro	10%	85%
incendiar, bombero, llama, sector, propagar	fuego	5%	80%

Cuadro 34: Reglas de asociación detectadas dentro de la temática incendios
Fuente: Elaboración propia

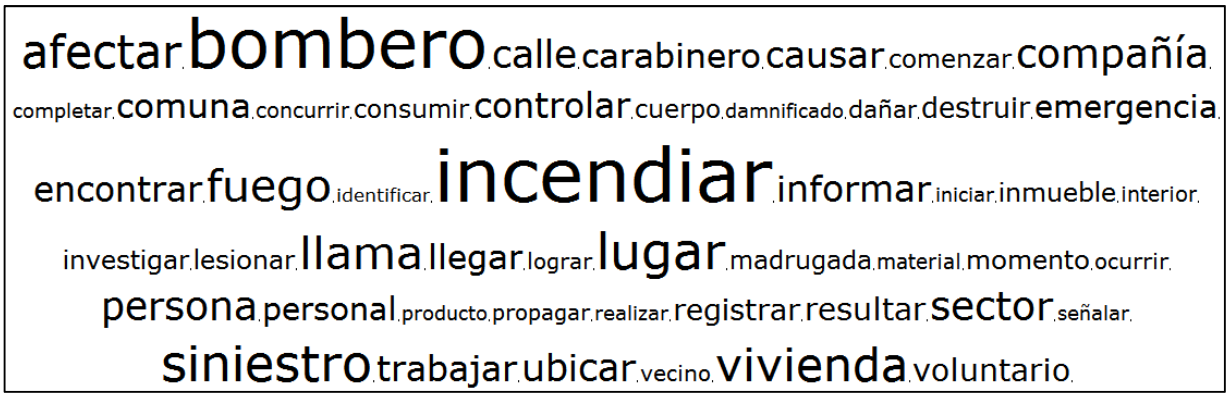

Figura 14: Tag cloud para la temática incendios
Fuente: Elaboración propia

4.3. Evaluación de la capacidad predictiva del modelo de identificación de temáticas policiales

A partir de los resultados obtenidos a través de la aplicación de clustering sobre las noticias policiales se estudia la capacidad de predicción de un modelo de clasificación de temáticas policiales, para lo cual se asume que cada cluster formado da origen a una clase. De esta forma se cuenta con un total de 14.045 noticias policiales distribuidas en siete clases distintas, sobre las cuales se aplica primeramente la lista de stop words, el stemming y una selección supervisada de atributos. Se establecen dos mecanismos para estudiar los modelos de clasificación, primero utilizando las 700 y luego las 1400 palabras con mayores valores para Chi-Cuadrado e Information Gain. Los tipos de modelos estudiados son Naive Bayes y K-nn (estudiados para distintos valores de K).

Chi-Cuadrado	
incendiar	delincuente
accidente	abusar
robar	marihuana
droga	homicidio
siniestro	colisionar
bombero	conductor
sexual	morir
cocaína	gramo
incautar	estudiante
llama	kilo

Information Gain	
robar	homicidio
accidente	fallecer
incendiar	tránsito
droga	conductor
bombero	detener
delincuente	cocaína
morir	vehículo
incautar	sujeto
siniestro	estudiante
colisionar	sexual

Cuadro 35: Palabras con los más altos valores de Chi-cuadrado e Information Gain. Fuente: Elaboración propia

Los distintos modelos son evaluados utilizando k-fold cross validation con k=10. Los modelos Naive Bayes mostraron un bajo desempeño para los esquemas de 700 y 1400 palabras con un accuracy de 71,12% y 68,20% respectivamente. Por otro lado, los modelos K-nn mostraron resultados bastantes similares entre los esquemas de 700 y 1.400 palabras, siendo levemente superior el esquema con 700 palabras, cuyo mejor desempeño se muestra para k=25 con un accuracy de 87,33%, precisión 88,15%, recall 85,74% y F-measure 86,76% (calculados de forma macro-averaged). Las medidas de desempeño del modelo K-nn para distintos valores de k se puede observar en la Figura 15 y Figura 16.

Para el modelo de clasificación seleccionado (k-nn con k=25 usando 700 palabras), las temáticas que mostraron mejor capacidad predictiva fueron drogas, robos, tránsito e incendios (con valores de F-measure bastantes similares). La temática delitos sexuales

muestra el desempeño más bajo con un recall de 74,77% y un F-measure de 78,83%. El detalle de las medidas de desempeño por temática se observa en el Cuadro 36.

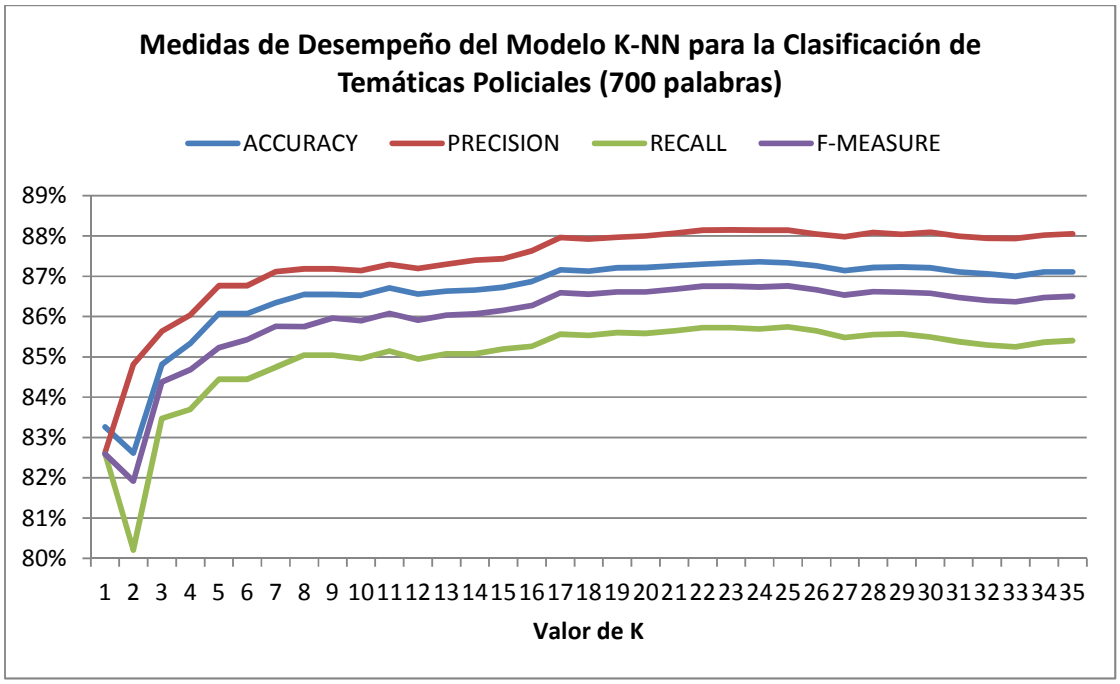

Figura 15: Medidas de desempeño modelo K-nn (con 700 palabras)
Fuente: Elaboración propia

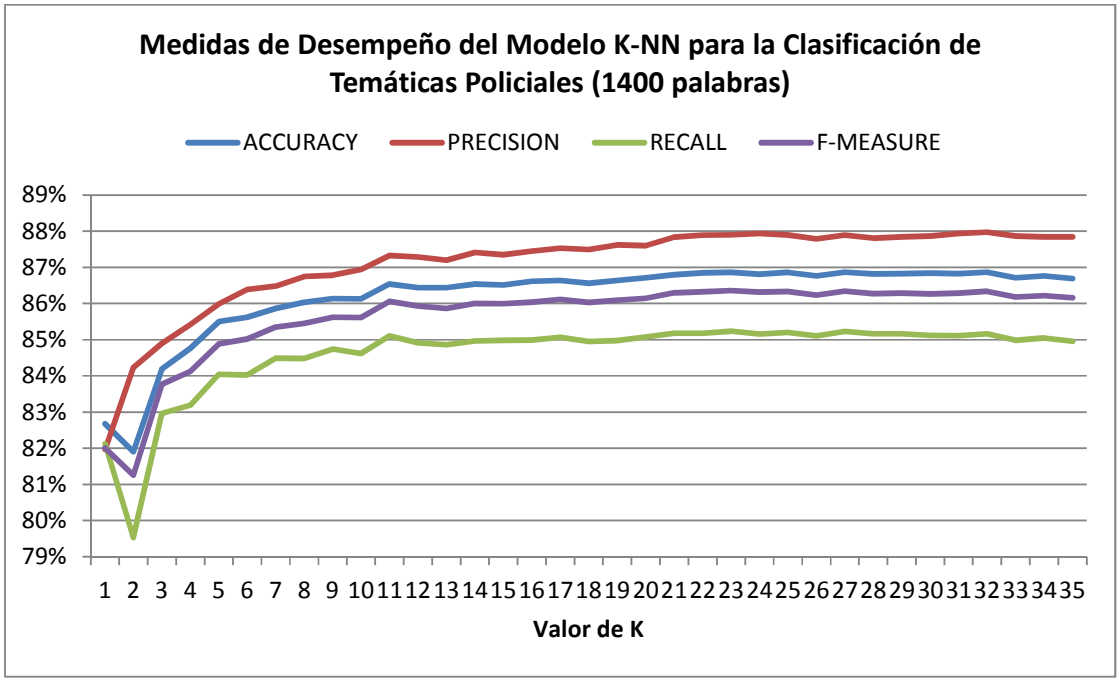

Figura 16: Medidas de desempeño modelo K-nn (con 1400 palabras)
Fuente: Elaboración propia

Temática	Precisión	Recall	F-measure
Drogas	94,39%	85,64%	89,80%
Robos	87,95%	93,95%	90,85%
Delitos Sexuales	83,35%	74,77%	78,83%
Homicidios	80,73%	88,57%	84,47%
Tránsito	88,01%	92,77%	90,33%
Disturbios	87,29%	80,05%	83,51%
Incendios	95,30%	84,46%	89,55%

Cuadro 36: Resumen medidas desempeño modelo k-nn (k=25) con 700 palabras
Fuente: Elaboración propia

4.4. Estadísticas de noticias policiales según región

En esta sección se estudian las estadísticas de las noticias policiales por región. Para determinar la región a la que pertenece cada noticia se explora en su contenido buscando la mención de alguna localidad asociada a cada región, por lo que para fines de este estudio una noticia está asociada a una o más regiones (un 78,5% se asocia a una región, un 17,3% a dos regiones y 4,2% a tres o más regiones). Se describe el número de noticias y el número de noticias por habitantes[6] según región (Figura 17 y Figura 18) y la distribución de las distintas temáticas dentro de cada región.

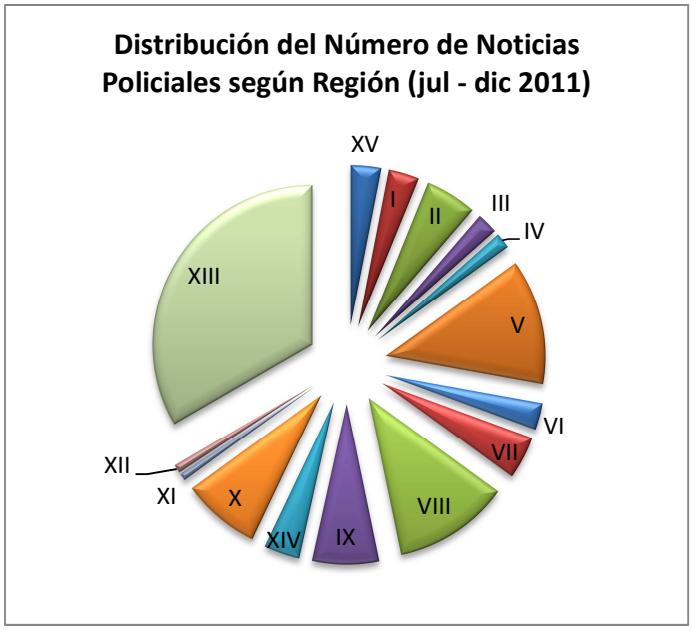

Figura 17: Distribución del n° de noticias policiales según región
Fuente: Elaboración propia

[6] Número de habitantes expresado en cientos de miles de acuerdo a la población estimada al 30 de junio de 2011 (basado en el censo 2002), disponible en: http://www.ine.cl/canales/menu/publicaciones/compendio_estadistico/pdf/2011/1.2demograficas.pdf

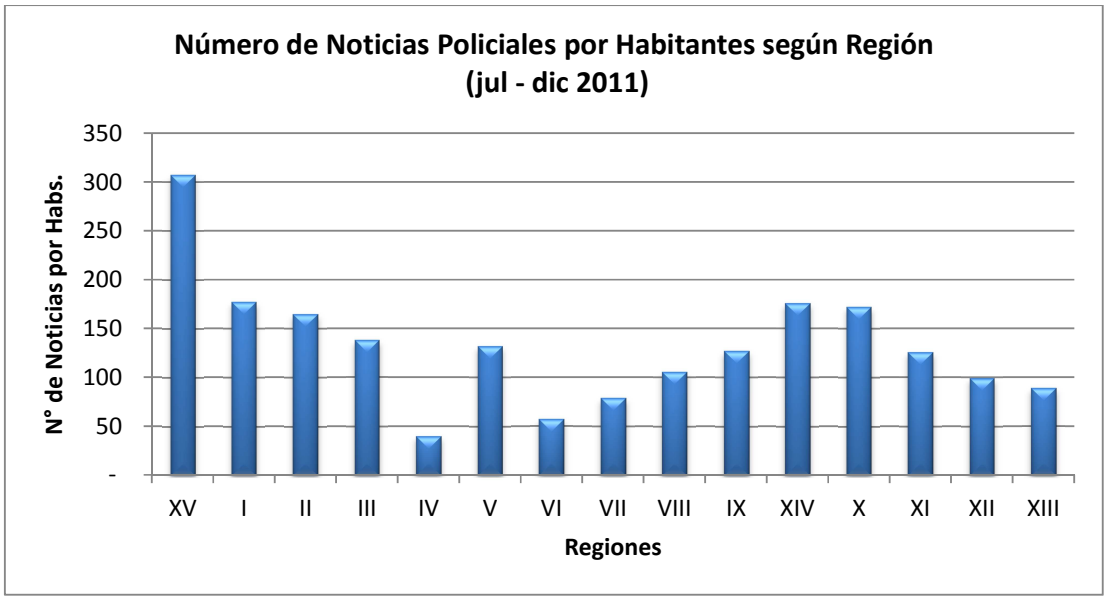

Figura 18: Número de noticias policiales por habitantes según región
Fuente: Elaboración propia

Se observa una correspondencia entre la posición relativa de las regiones ordenadas de acuerdo al número de noticias y la posición relativa de las regiones ordenadas de acuerdo al número de habitantes (Figura 19): en general a mayor número de habitantes de una región, mayor número de noticias asociadas a esta. Destaca la situación de regiones en que sería esperable una mayor cantidad de noticias policiales dado su población como es el caso de las regiones de Coquimbo, del Lib. Bernardo O'Higgins y del Maule; mientras que la situación contraria se da en las regiones de Arica y Parinacota, de Antofagasta y de Los Lagos.

Estudiando la posición relativa de las regiones respecto al número de noticias por habitantes se determina que no se existe una relación con la posición relativa de acuerdo al número de noticias (Figura 20). De esta forma, regiones como la Metropolitana, que posee por lejos los más altos niveles de noticias, se transforma en una de las regiones con menor índice noticias por habitantes. Por otro lado, destaca los altos niveles de noticias por habitantes que alcanza la región de Arica y Parinacota, derivado de un relativo alto número de noticias policiales para una población pequeña.

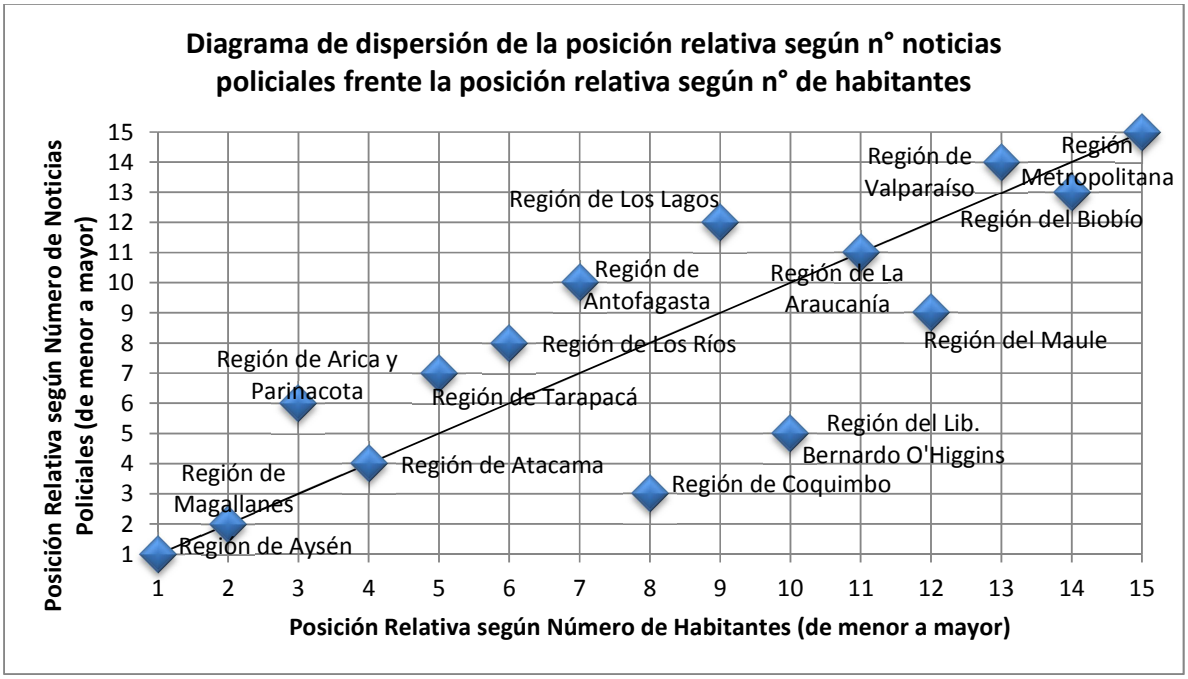

Figura 19: Diagrama de dispersión de la posición relativa según n° noticias policiales frente la posición relativa según n° de habitantes
Fuente: Elaboración propia

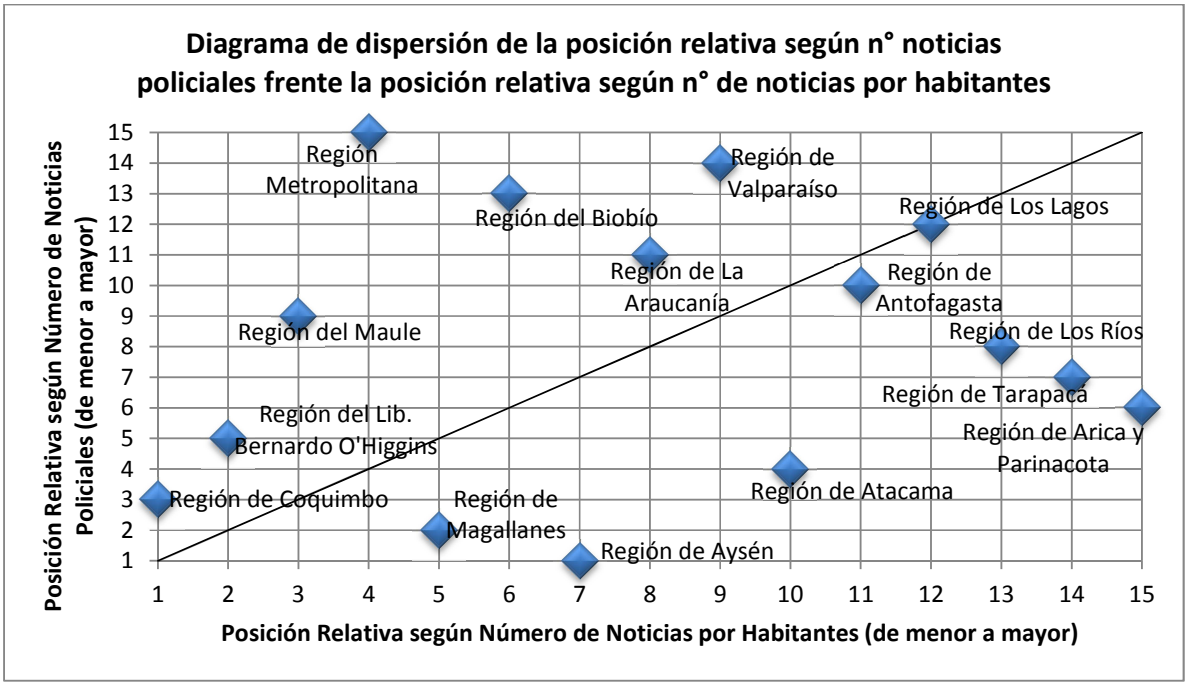

Figura 20: Diagrama de dispersión de la posición relativa según n° noticias policiales frente la posición relativa según n° de noticias por habitantes
Fuente: Elaboración propia

A partir de los resultados obtenidos en la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC 2011), mostrados en la Figura 21 y Figura 22, es posible apreciar que aquellas regiones que muestran una menor diferencia entre el porcentaje de personas que cree que será víctima durante el 2012 y el porcentaje de hogares victimizados el 2011 corresponden efectivamente a las regiones que presentan los niveles de noticias policiales por habitantes más bajos (regiones IV, VI, VII y XIII). Así mismo, dentro de las regiones que muestran una mayor diferencia entre ambos porcentajes también muestran un relativo mayor nivel de noticias policiales por habitantes (regiones XV y II).

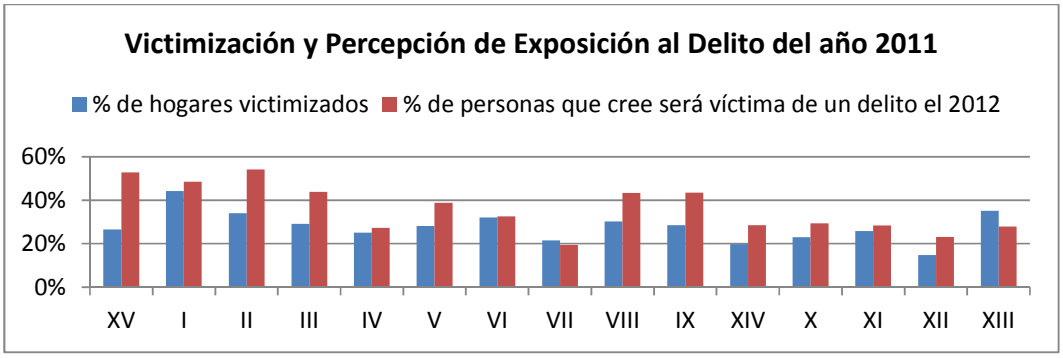

Figura 21: Victimización y percepción de exposición al delito del año 2011
Fuente: Elaboración propia

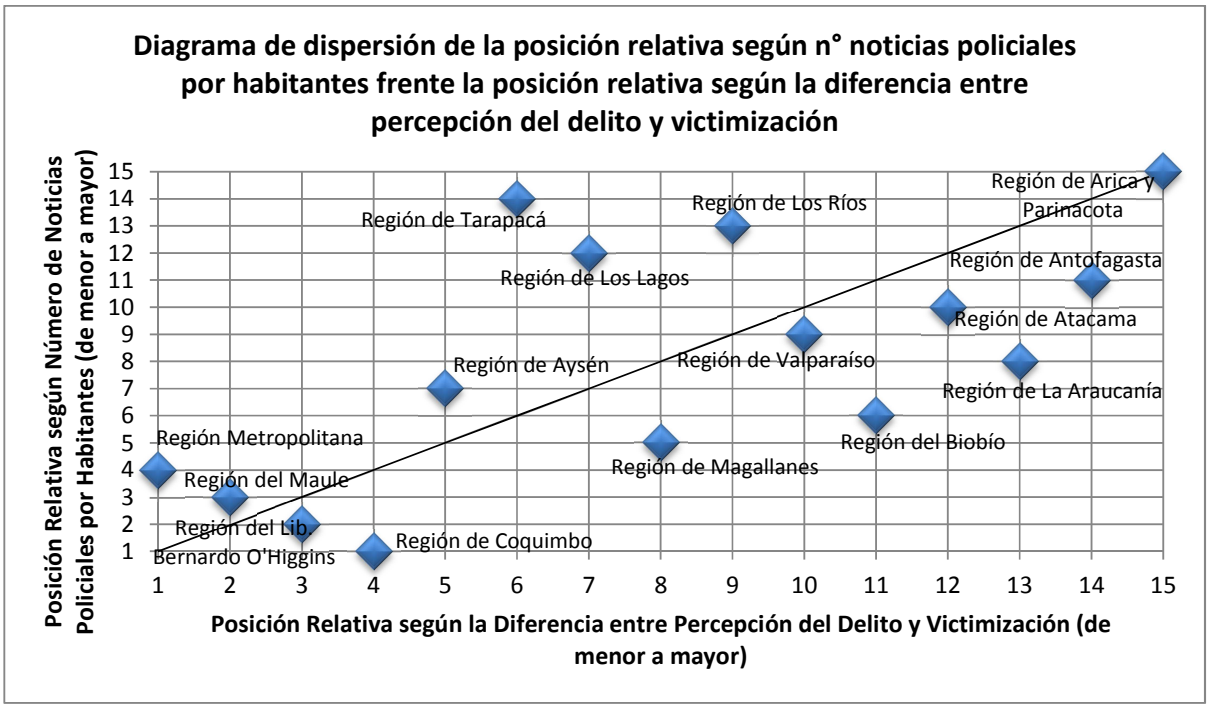

Figura 22: Diagrama de dispersión del n° noticias policiales por habitantes frente la diferencia entre percepción del delito y victimización
Fuente: Elaboración propia

A continuación se presenta la distribución de las distintas temáticas dentro de las noticias policiales para cada una de las regiones del país durante el período de estudio:

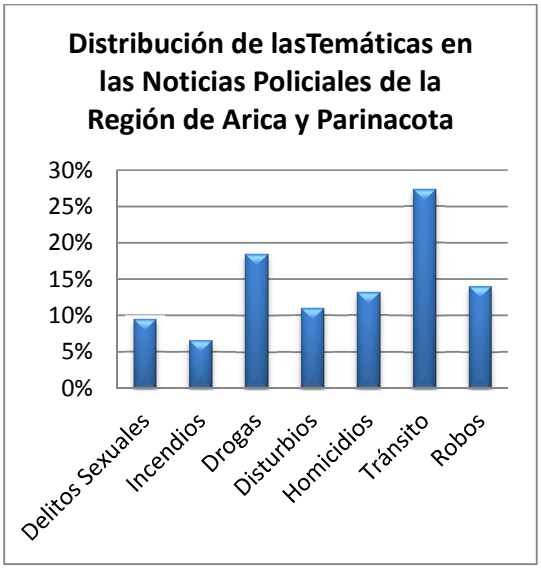

Figura 23: Distribución de las temáticas en la XV Región
Fuente: Elaboración propia

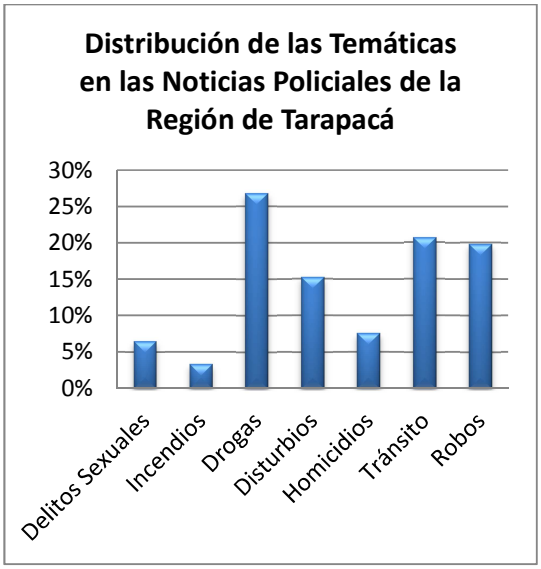

Figura 24: Distribución de las temáticas en la I Región
Fuente: Elaboración propia

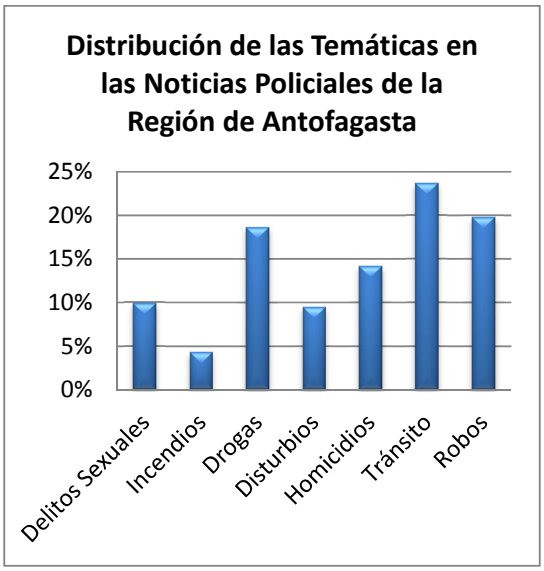

Figura 25: Distribución de las temáticas en la II Región
Fuente: Elaboración propia

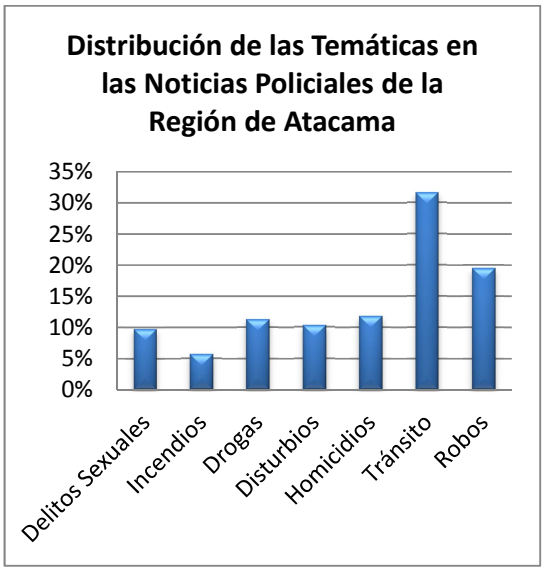

Figura 26: Distribución de las temáticas en la III Región
Fuente: Elaboración propia

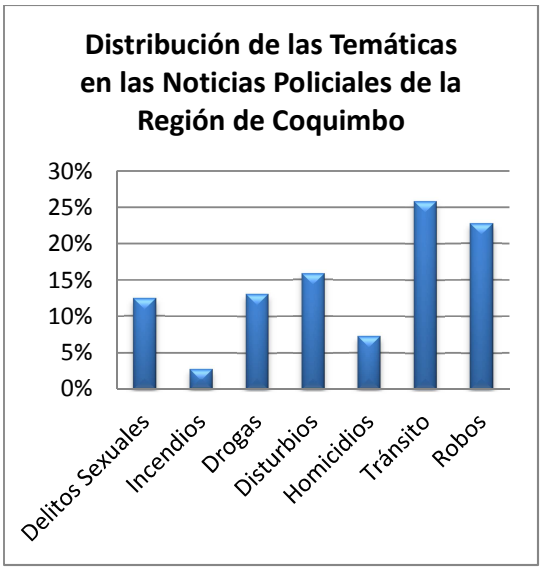

Figura 27: Distribución de las temáticas en la IV Región
Fuente: Elaboración propia

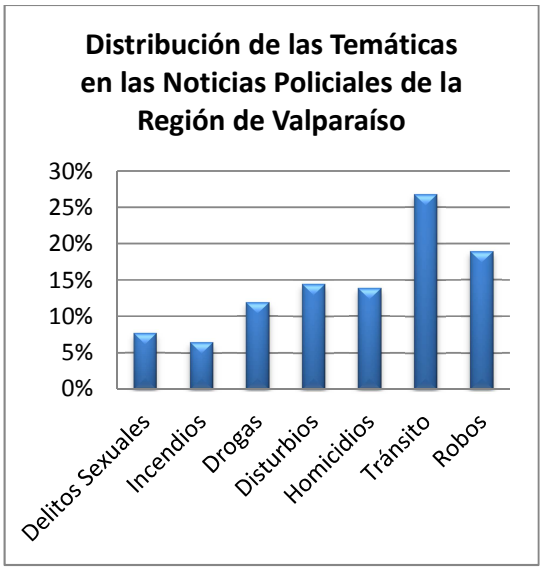

Figura 28: Distribución de las temáticas en la V Región
Fuente: Elaboración propia

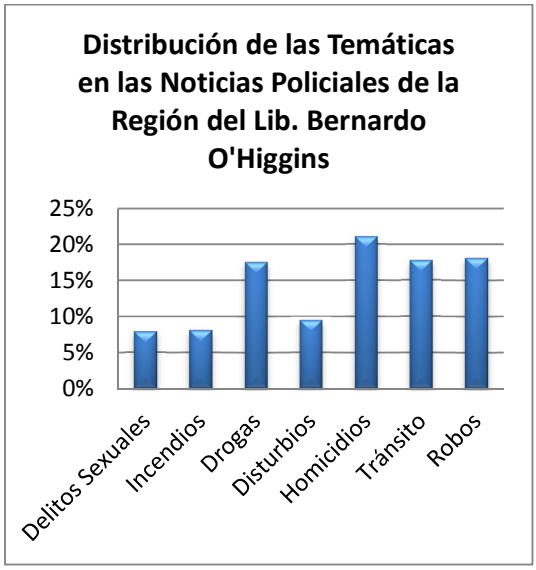

Figura 29: Distribución de las temáticas en la VI Región
Fuente: Elaboración propia

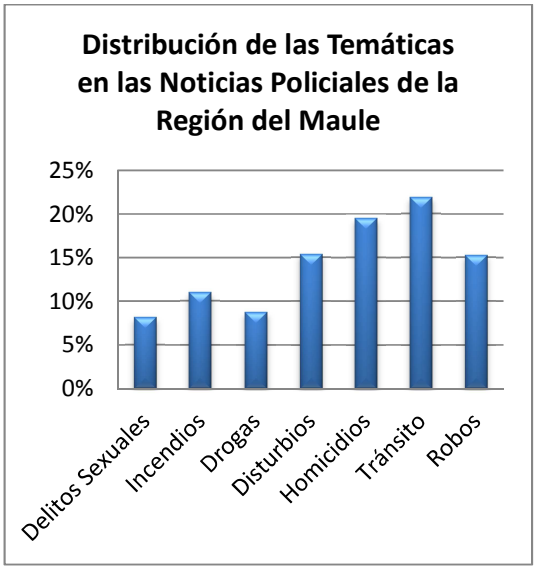

Figura 30: Distribución de las temáticas en la VII Región
Fuente: Elaboración propia

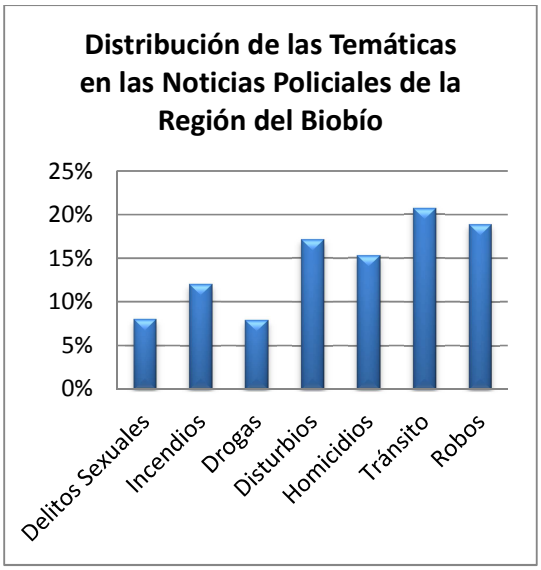

Figura 31: Distribución de las temáticas en la VIII Región
Fuente: Elaboración propia

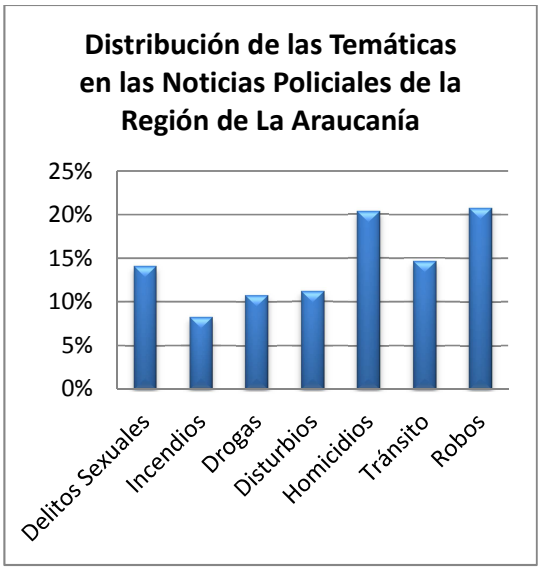

Figura 32: Distribución de las temáticas en la IX Región
Fuente: Elaboración propia

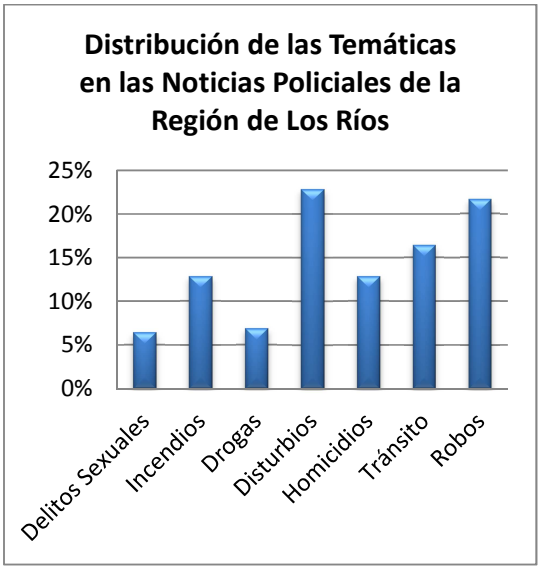

Figura 33: Distribución de las temáticas en la XIV Región
Fuente: Elaboración propia

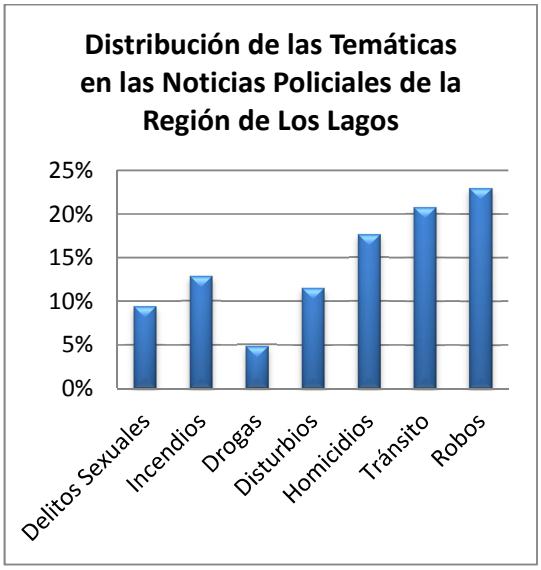

Figura 34: Distribución de las temáticas en la X Región
Fuente: Elaboración propia

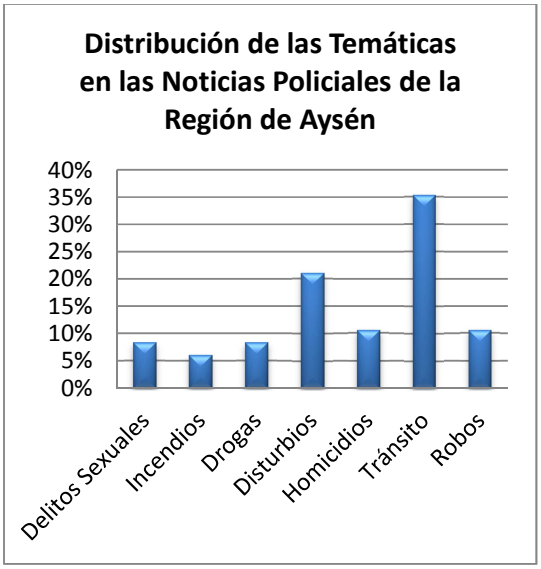

Figura 35: Distribución de las temáticas en la XI Región
Fuente: Elaboración propia

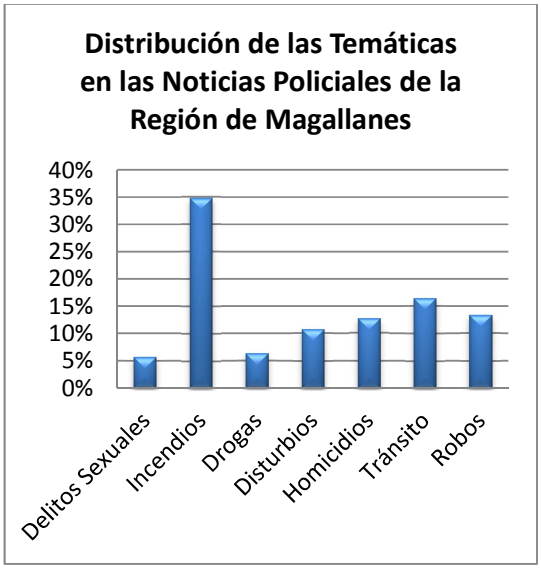

Figura 36: Distribución de las temáticas en la XII Región
Fuente: Elaboración propia

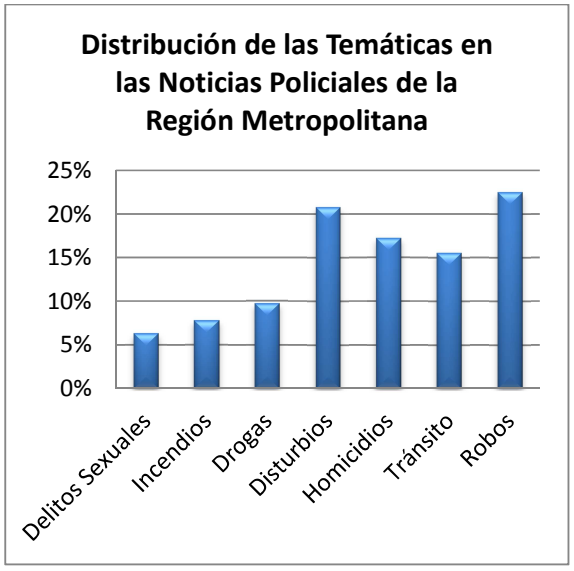

Figura 37: Distribución de las temáticas en la XIII Región
Fuente: Elaboración propia

Entre las temáticas con mayor cobertura durante el período de estudio se encuentran: tránsito (para 8 regiones) y robos (para 3 regiones). Destaca el caso de la I región en la cual la mayor cobertura es para la temática drogas.

Entre las temáticas con menor cobertura durante período de estudio se encuentran: incendios (para 8 regiones), delitos sexuales (para 5 regiones) y drogas (para 2 regiones).

4.5. Estadísticas comparativas sobre noticias y casos reales por habitantes según temática policial.

4.5.1. Temática drogas

A partir de la Figura 38 y Figura 39 se detecta la tendencia de un alto número de casos y de noticias por habitantes sobre drogas en las regiones XV, I y II, mientras que para el resto de las regiones tanto los niveles de casos como de noticias por habitantes se reducen notoriamente. Se observa que en las regiones IV, VII, VIII, IX, XIV, X, XI y XII los niveles de casos por habitantes son relativamente muy bajos con respecto al resto de las regiones, manteniéndose relativamente constantes los niveles de noticias por habitantes entre estas regiones. A nivel general se mantiene una proporcionalidad entre el número de noticias y el número de casos por habitantes para las distintas regiones (Figura 40), pero habría sido esperable tener un menor número de noticias en las regiones XV, II, III, V, IX y XIV.

Figura 38: Nivel de noticias por habs. sobre drogas, según región
Fuente: Elaboración propia

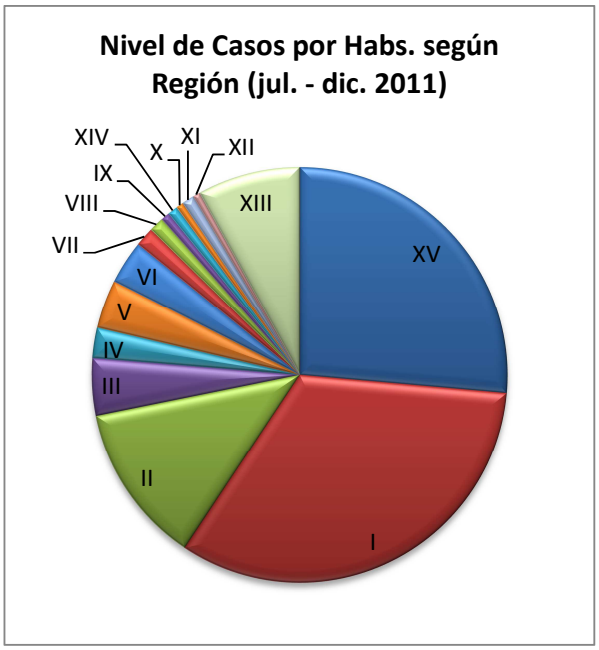

Figura 39: Nivel de casos por habs. sobre drogas[7], según región
Fuente: Elaboración propia

[7] El número de casos policiales sobre Ley de Drogas se obtiene a través de la solicitud de información pública a Carabineros de Chile (Ley 20285 de Transparencia)

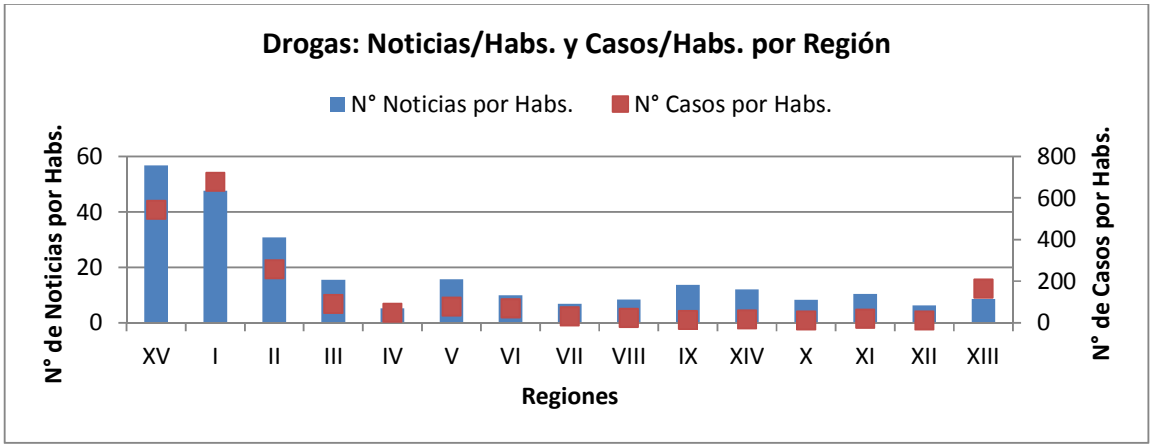

Figura 40: Noticias/habs. y casos/habs. sobre drogas, según región
Fuente: Elaboración propia

4.5.2. Temática robos

De acuerdo a la Figura 43 se observa una relativa buena proporcionalidad entre el número de casos y de noticias por habitantes para varias de las regiones (I, II, III, V, VII, VIII XI y XII). En regiones como la XV, XIV y X se habría esperado un menor número de noticias, mientras que la situación contraria se habría esperado para regiones como la IV, VI y XIII. En general se observa una mayor variabilidad en los niveles de noticias por habitantes que de los niveles de casos por habitantes entre las distintas regiones durante el período de estudio (Figura 41 y Figura 42).

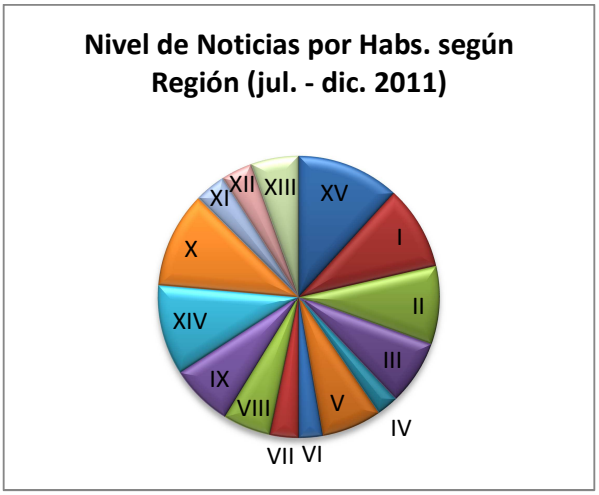

Figura 41: Nivel de noticias por habs. sobre robos, según región
Fuente: Elaboración propia

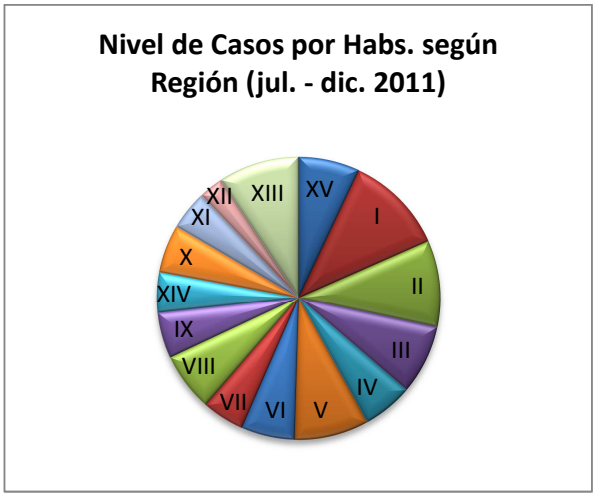

Figura 42: Nivel de casos por habs. sobre robos[8], según región
Fuente: Elaboración propia

[8] El número de casos policiales sobre Robos se obtiene a través información disponible en el sitio web de la Subsecretaría de Prevencion del Delito del Gobierno de Chile: http://www.seguridadpublica.gov.cl/delitos_de_mayor_connotacion_social.html

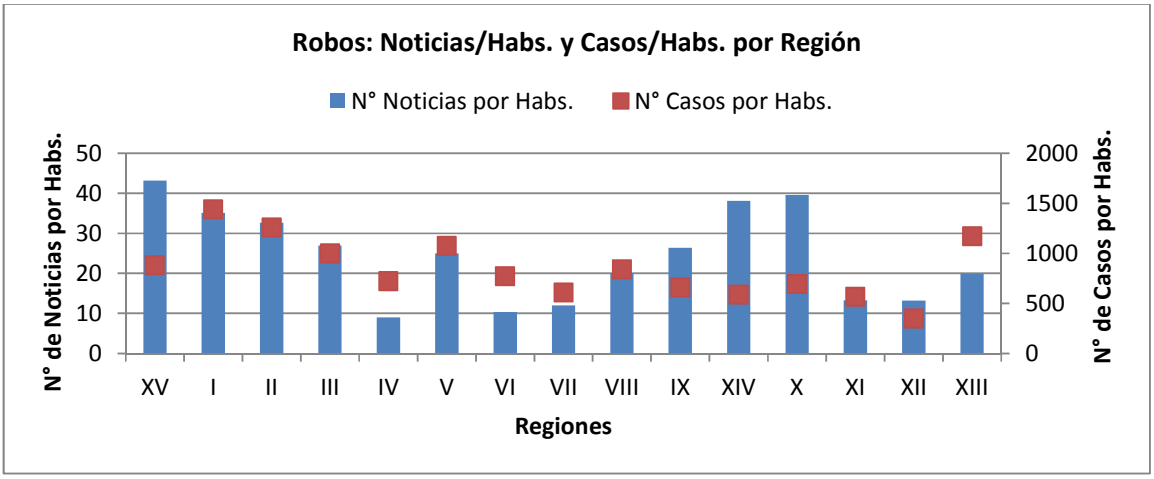

Figura 43: Noticias/habs. y casos/habs. sobre robos, según región
Fuente: Elaboración propia

4.5.3. Temática delitos sexuales

A partir de la Figura 44 y Figura 45 se puede constatar que el número de casos por habitantes se mantiene relativamente constante para las distintas regiones, mientras que el número de noticias por habitantes muestra un claro contraste entre las regiones. Entre las regiones que se observa un mayor número de noticias por habitantes se encuentran la XV, II, IX y X. En la Figura 46 no se observa una proporcionalidad entre los niveles de casos y de noticias por habitantes durante el período de estudio.

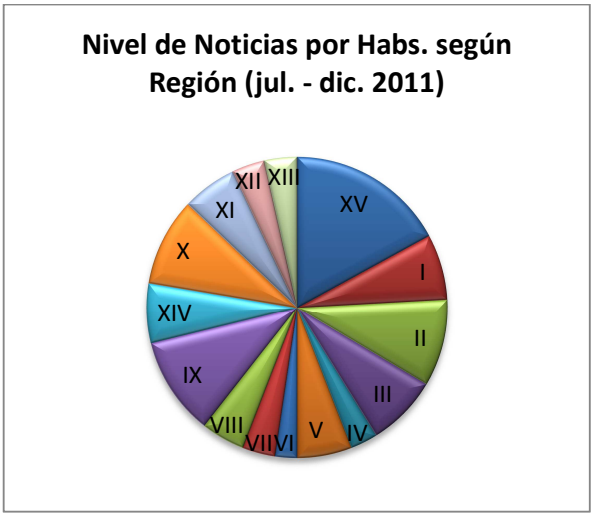

Figura 44: Nivel de noticias por habs. sobre delitos sexuales, según región
Fuente: Elaboración propia

Figura 45: Nivel de casos por habs. sobre delitos sexuales[9], según región
Fuente: Elaboración propia

[9] El número de casos policiales sobre Delitos de Connotación Sexual se obtiene a través de la solicitud de información pública a Carabineros de Chile (Ley 20285 de Transparencia)

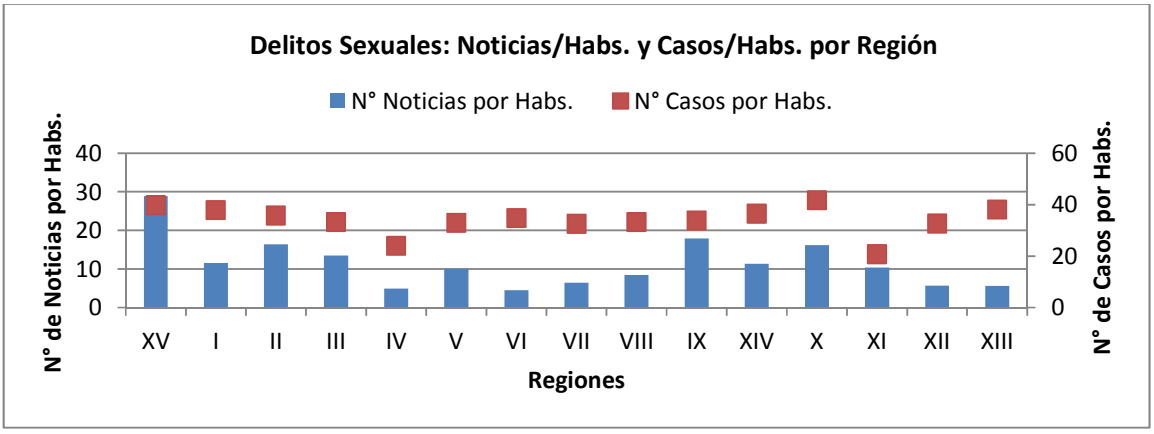

Figura 46: Noticias/habs. y casos/habs. sobre delitos sexuales, según región
Fuente: Elaboración propia

4.5.4. Temática homicidios

De acuerdo a la Figura 49 se aprecia que el número de casos de homicidios por habitantes se mantiene relativamente bajo y constante para la mayoría regiones, mientras que los niveles de noticias por habitantes presentan una mayor variabilidad (Figura 47). Destacan los casos de la región XV y X que presentan los más altos niveles de noticias por habitantes. Solo para las regiones IV, VI XI se esperaría un considerable mayor nivel de noticias.

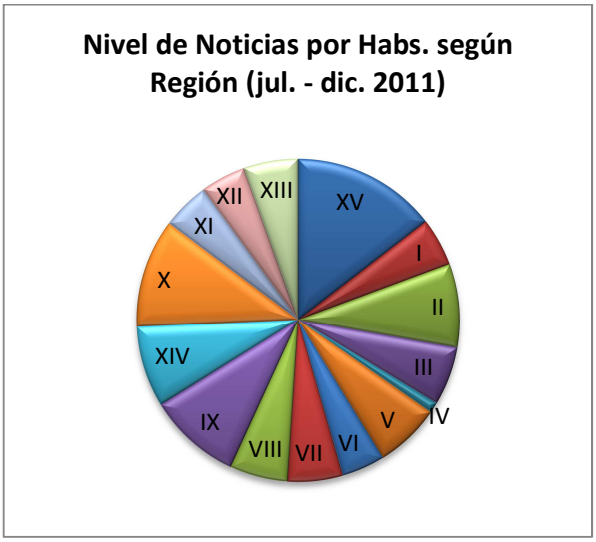

Figura 47: Nivel de noticias por habs. sobre homicidios, según región
Fuente: Elaboración propia

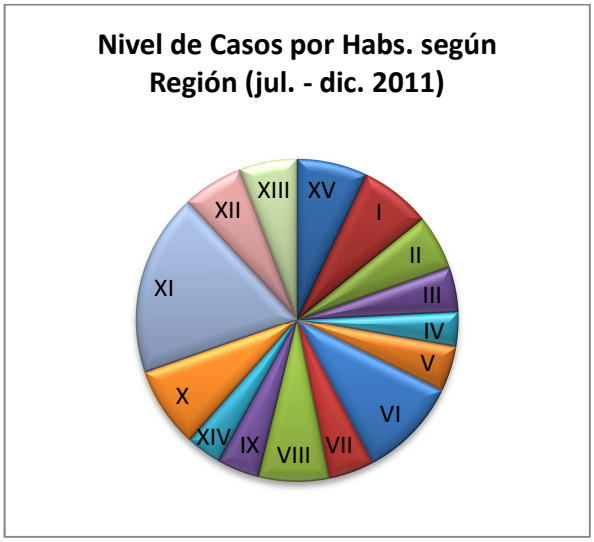

Figura 48: Nivel de casos por habs. sobre homicidios[10], según región
Fuente: Elaboración propia

[10] El número de casos policiales sobre Homicidios se obtiene a través información disponible en el sitio web de la Subsecretaría de Prevencion del Delito del Gobierno de Chile: http://www.seguridadpublica.gov.cl/delitos_de_mayor_connotacion_social.html

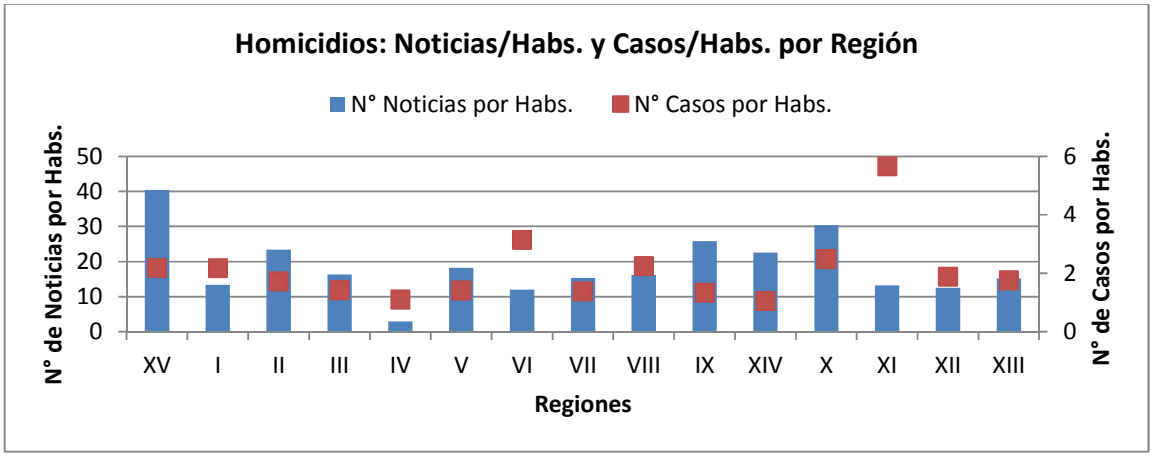

Figura 49: Noticias/habs. y casos/habs. sobre homicidios, según región
Fuente: Elaboración propia

4.5.5. Temática tránsito

A partir de la Figura 52 se observa una relativa buena proporcionalidad entre los niveles de casos y de noticias por habitantes para las regiones que presentan un alto nivel de noticias por habitantes, como es el caso de las regiones XV, I, II, III, XIV y X. Los niveles de casos por habitantes se presentan relativamente constantes para la mayoría de las regiones (Figura 51), mientras que el número de noticias por habitantes según región experimenta una mayor variabilidad (Figura 50).

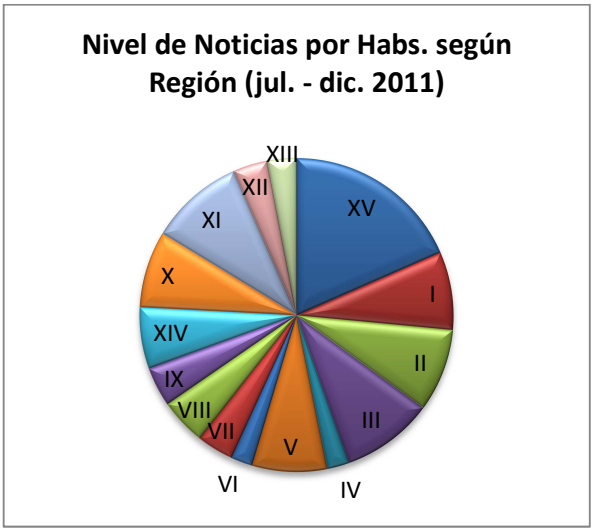

Figura 50: Nivel de noticias por habs. sobre tránsito, según región
Fuente: Elaboración propia

Figura 51: Nivel de casos por habs. sobre tránsito[11], según región
Fuente: Elaboración propia

[11] El número de casos policiales sobre Accidentes de Tránsito se obtiene a través de la solicitud de información pública a Carabineros de Chile (Ley 20285 de Transparencia)

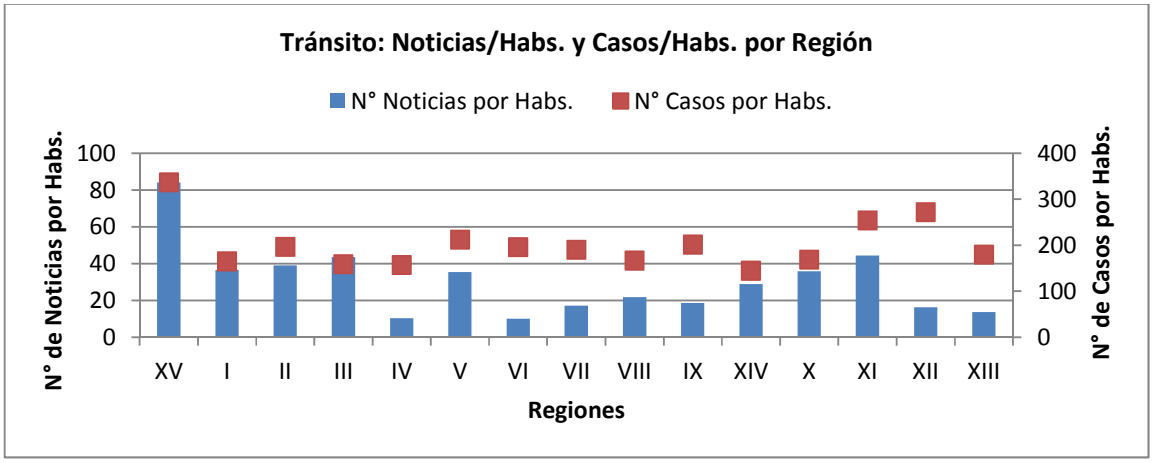

Figura 52: Noticias/habs. y casos/habs. sobre tránsito, según región
Fuente: Elaboración propia

4.5.6. Temática disturbios

En la Figura 55 no se aprecia una proporcionalidad entre los niveles de casos y de noticias por habitantes según región. En varias regiones se esperaría tener un mayor nivel de noticias (regiones II, III, V, VI, XII y XIII) asumiendo una relación entre ambas variables en estudio, pero se observa un relativo buen ajuste entre las variables en regiones como la XV, IV, VIII, XIV y X. Tanto los niveles de casos como de noticias por habitantes presentan una relativa alta variabilidad entre las diferentes regiones (Figura 53 y Figura 54).

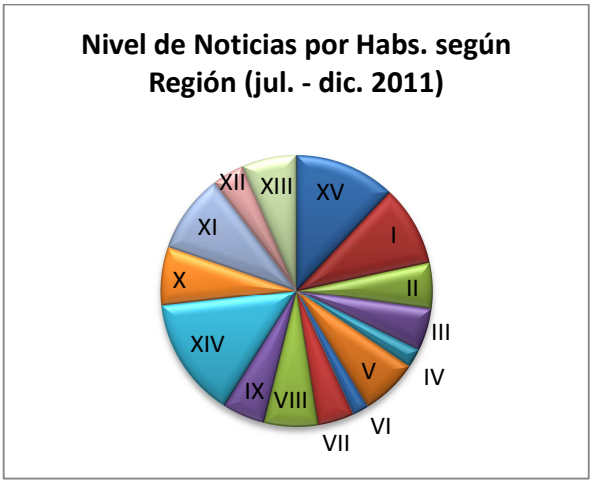

Figura 53: Nivel de noticias por habs. sobre disturbios, según región
Fuente: Elaboración propia

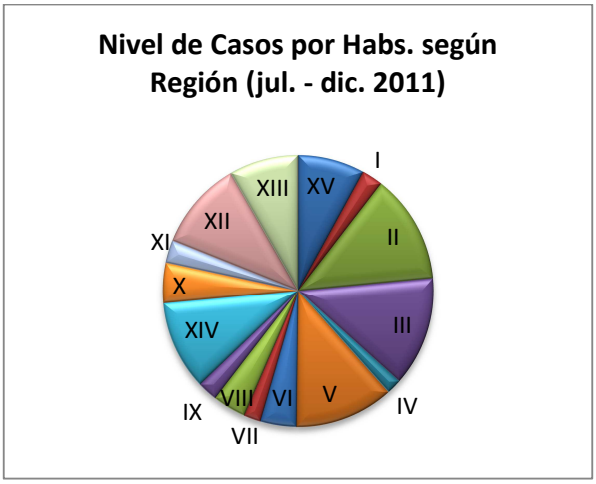

Figura 54: Nivel de casos por habs. sobre disturbios[12], según región
Fuente: Elaboración propia

[12] El número de casos policiales sobre Desórdenes Públicos se obtiene a través de la solicitud de información pública a Carabineros de Chile (Ley 20285 de Transparencia)

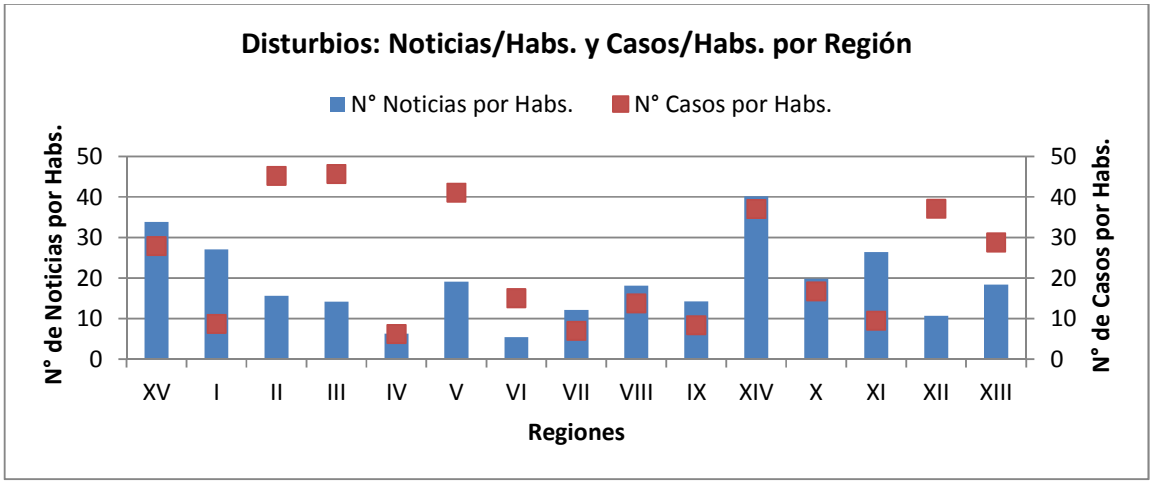

Figura 55: Noticias/habs. y casos/habs. sobre disturbios, según región
Fuente: Elaboración propia

4.5.7. Temática incendios

En la Figura 57 se observa que los niveles de casos por habitantes se mantienen constantes para la mayoría de las regiones, destacando por tener niveles distintivamente altos las regiones XIV, XI y XII. Se presenta una relativa buena proporcionalidad entre los niveles de noticias y casos por habitantes (Figura 58).

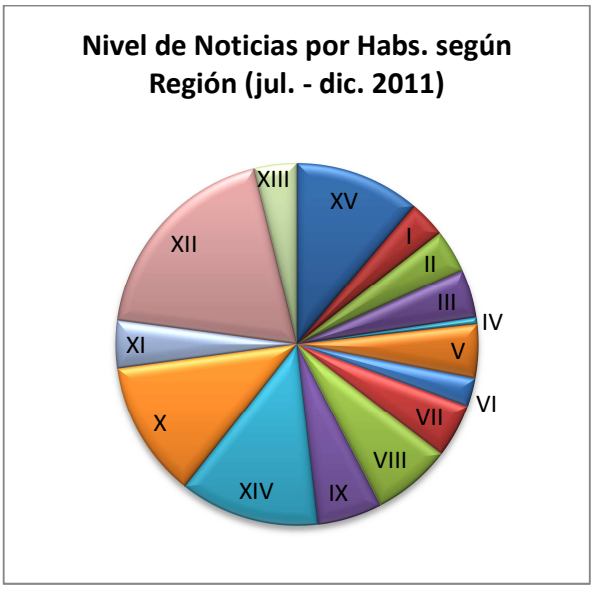

Figura 56: Nivel de noticias por habs. sobre incendios, según región
Fuente: Elaboración propia

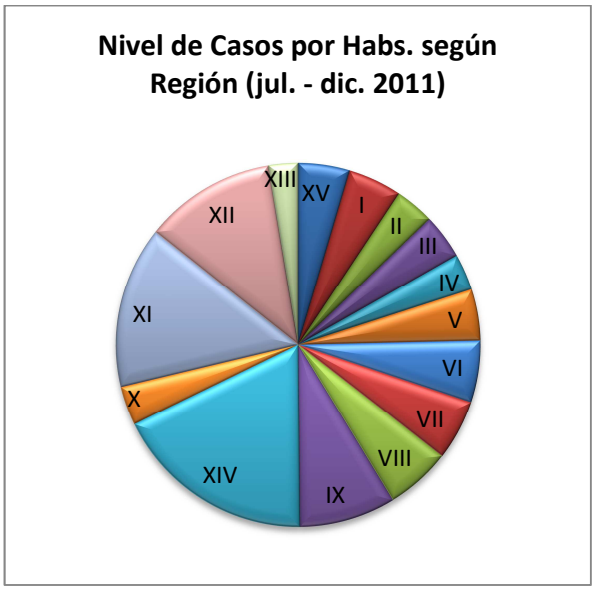

Figura 57: Nivel de casos por habs. sobre incendios[13], según región
Fuente: Elaboración propia

[13] El número de casos sobre Incendios se obtiene a través de la solicitud de información pública a Carabineros de Chile (Ley 20285 de Transparencia)

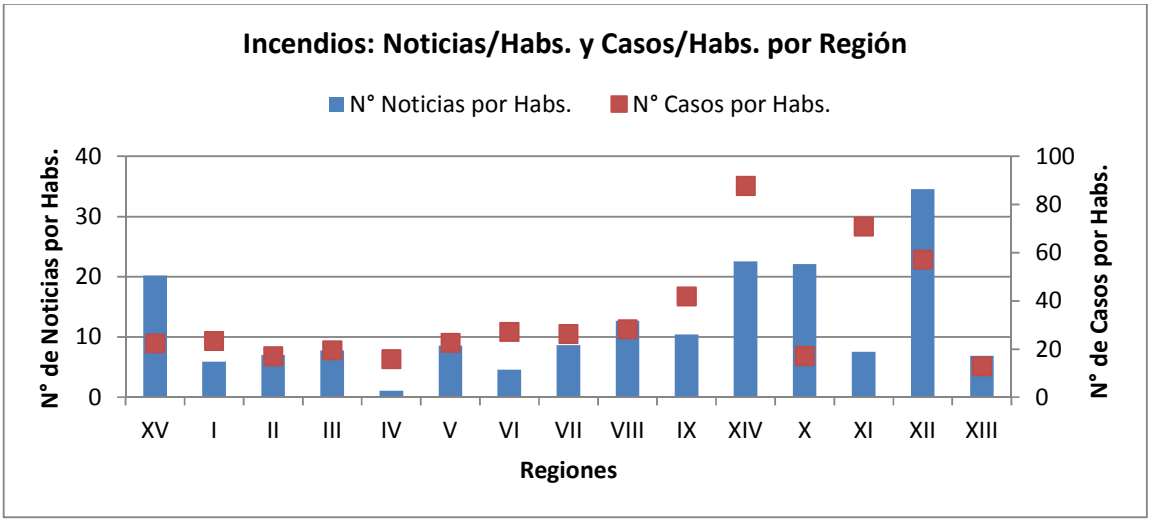

Figura 58: Noticias/habs. y casos/habs. sobre incendios, según región
Fuente: Elaboración propia

4.6. Análisis de regresiones lineales simples

En esta sección se estudia las posibles relaciones lineales simples que puedan existir entre el número de casos y el número de noticias policiales mensuales para las distintas temáticas utilizando los datos aportados por cada región, logrando así reunir 90 pares de datos para cada temática (Anexo 7.1.). Se pretende encontrar un modelo lineal en los parámetros que describa una correspondencia entre el número de casos y de noticias, evaluando para ello distintas transformaciones sobre ambas variables (en particular la función raíz cuadrada y la función logaritmo) que cumplan con los supuestos básicos de un modelo de regresión lineal (Anexo 7.2. y Apéndice 8.1.).

4.6.1. Temática drogas

La Figura 59 muestra, para la temática drogas, una aparente relación lineal entre la raíz cuadrada del número de casos y la raíz cuadrada del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$\sqrt{(N° \, Noticias)} = 2{,}069 + 0{,}190 * \sqrt{(N° \, Casos)}$$

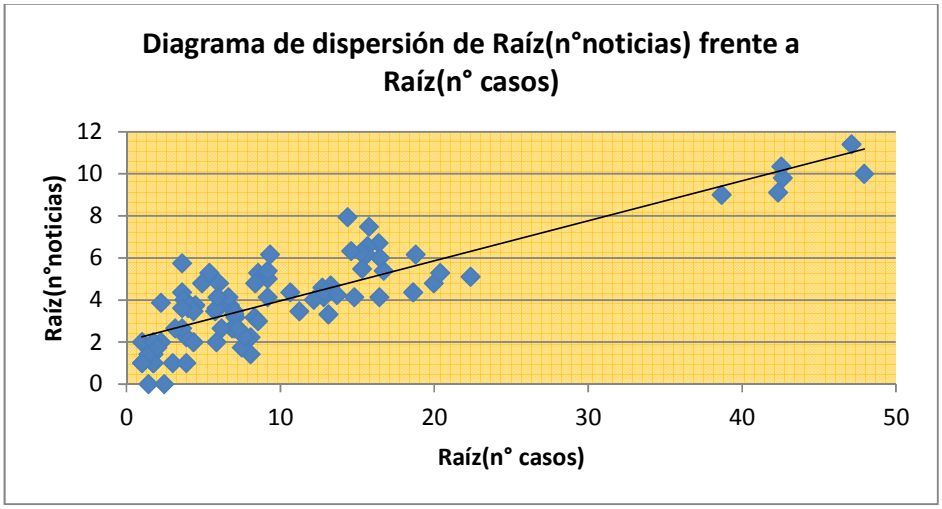

Figura 59: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre drogas
Fuente: Elaboración propia

R	R^2	Error Típico de la Estimación
0,851	0,723	1,216

Cuadro 37: Resumen del modelo de regresión lineal para la temática drogas
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una relación estadísticamente significativa entre la raíz del número de casos y la raíz del número de noticias (con un p-valor aproximadamente a cero), la cual explicaría el 72,3% de la variabilidad en la raíz del número de noticias.

Los datos referentes a la región metropolitana muestran valores relativamente más altos que en el resto de las regiones, por lo que podría tener una influencia demasiado importante en los resultados. El estudio sin considerar los datos aportados por la región metropolitana muestra un coeficiente de correlación de 0,707 entre las variables analizadas, por lo que se decide mantener el modelo inicial, pues se conserva relativamente fuerte la probable relación lineal entre las variables. En la Figura 60 se observa que la ausencia de los valores relacionados a la región metropolitana no modifica la ecuación de la regresión de manera sustancial.

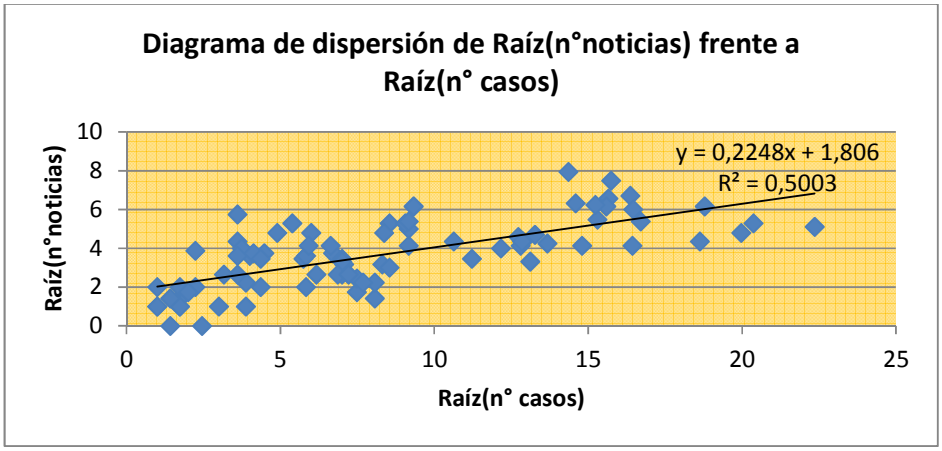

Figura 60: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre drogas sin los datos de la XIII región
Fuente: Elaboración propia

4.6.2. Temática robos

La Figura 61 muestra, para la temática robos, una aparente relación lineal entre la raíz cuadrada del número de casos y la raíz cuadrada del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

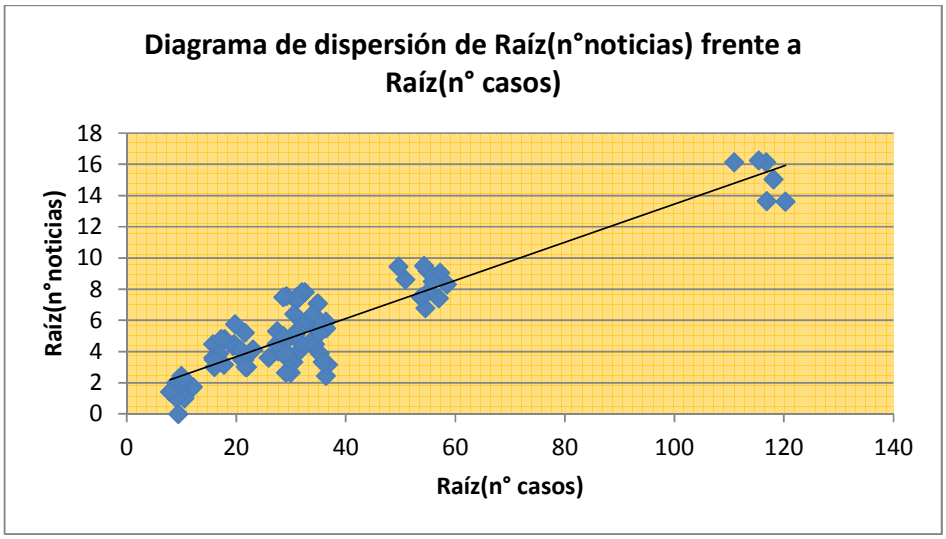

Figura 61: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre robos
Fuente: Elaboración propia

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$\sqrt{(N°\ Noticias)} = \beta_0 + \beta_1 * \sqrt{(N°\ Casos)} + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$\sqrt{(N° \, Noticias)} = 1{,}217 + 0{,}122 * \sqrt{(N° \, Casos)}$$

R	R^2	Error Típico de la Estimación
0,920	0,847	1,333

Cuadro 38: Resumen del modelo de regresión lineal para la temática robos
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una relación estadísticamente significativa entre la raíz del número de casos y la raíz del número de noticias (con un p-valor aproximadamente cero), la cual explicaría el 84,7% de la variabilidad en la raíz del número de noticias.

Los datos referentes a la región metropolitana muestran valores relativamente más altos que en el resto de las regiones, por lo que podría tener una influencia demasiado importante en los resultados. El estudio sin considerar los datos aportados por la región metropolitana muestra un coeficiente de correlación de 0,812 entre las variables analizadas, por lo que se decide mantener el modelo inicial, pues se conserva relativamente fuerte la probable relación lineal entre las variables. En la Figura 62 se observa que la ausencia de los valores relacionados a la región metropolitana no modifica la ecuación de la regresión de manera sustancial.

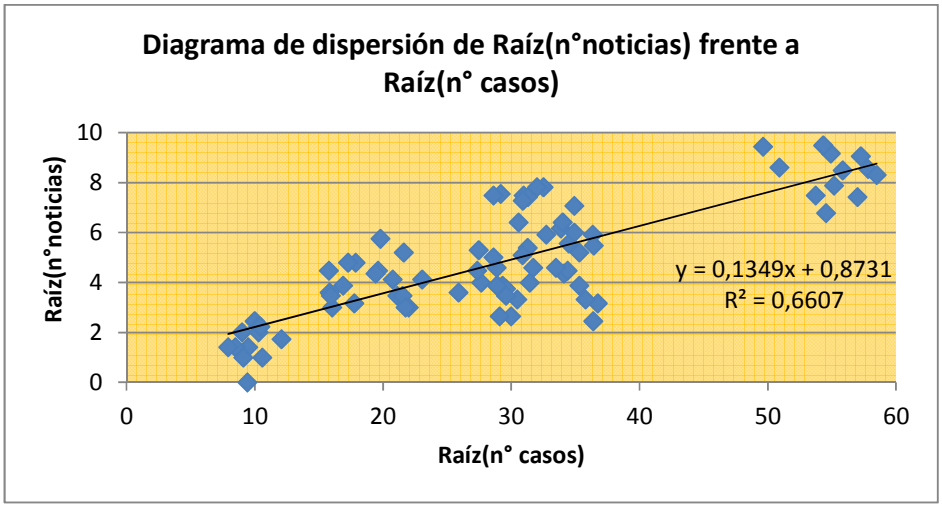

Figura 62: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre robos sin los datos de la XIII región
Fuente: Elaboración propia

4.6.3. Temática delitos sexuales

La Figura 63 muestra, para la temática delitos sexuales, una aparente relación lineal entre la raíz cuadrada del número de casos y la raíz cuadrada del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$\sqrt{(N° \, Noticias)} = 1{,}054 + 0{,}362 * \sqrt{(N° \, Casos)}$$

R	R^2	Error Típico de la Estimación
0,836	0,700	1,077

Cuadro 39: Resumen del modelo de regresión lineal para la temática delitos sexuales
Fuente: Elaboración propia

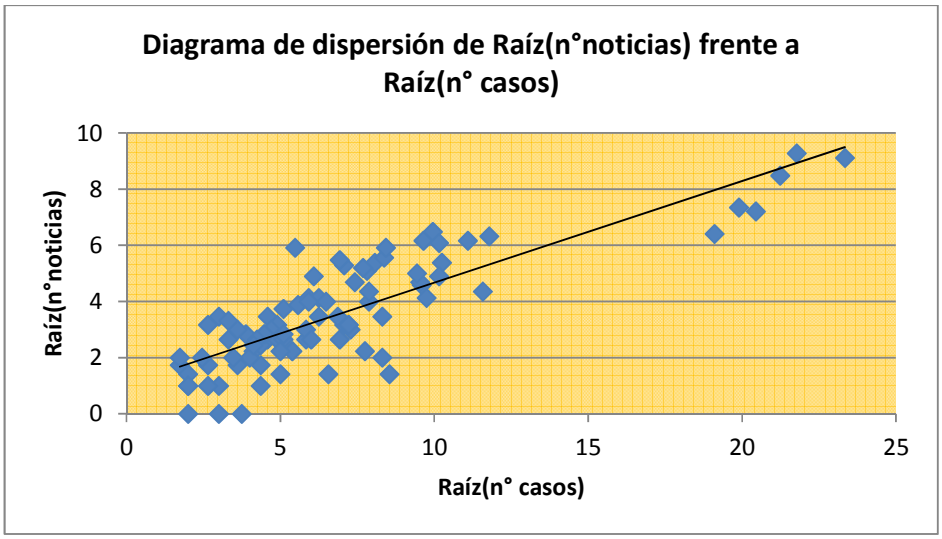

Figura 63: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre delitos sexuales
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una relación estadísticamente significativa entre la raíz del número de casos y la raíz del

número de noticias (con un p-valor aproximadamente a cero), la cual explicaría el 70,0% de la variabilidad en la raíz del número de noticias.

Los datos referentes a la región metropolitana muestran valores relativamente más altos que en el resto de las regiones, por lo que podría tener una influencia demasiado importante en los resultados. El estudio sin considerar los datos aportados por la región metropolitana muestra un coeficiente de correlación de 0,750 entre las variables analizadas, por lo que se decide mantener el modelo inicial, pues se conserva relativamente fuerte la probable relación lineal entre las variables. En la Figura 64 se observa que la ausencia de los valores relacionados a la región metropolitana no modifica la ecuación de la regresión de manera sustancial.

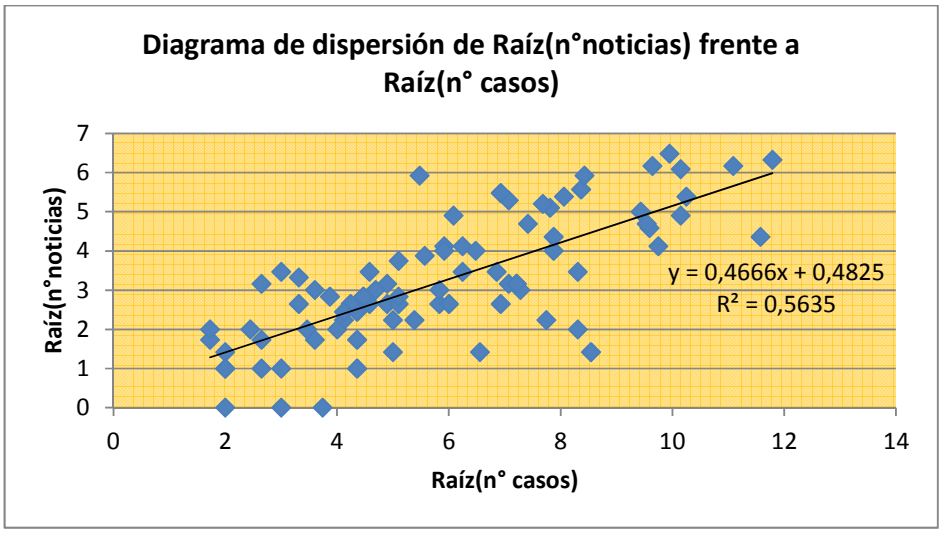

Figura 64: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre delitos sexuales sin los datos de la XIII región
Fuente: Elaboración propia

4.6.4. Temática homicidios

La Figura 65 muestra, para la temática homicidios, una aparente relación lineal entre la raíz cuadrada del número de casos y la raíz cuadrada del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$\sqrt{(N°\ Noticias)} = \beta_0 + \beta_1 * \sqrt{(N°\ Casos)} + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$\sqrt{(N°\ Noticias)} = 1{,}539 + 2{,}193 * \sqrt{(N°\ Casos)}$$

R	R^2	Error Típico de la Estimación
0,810	0,656	1,842

Cuadro 40: Resumen del modelo de regresión lineal para la temática homicidios
Fuente: Elaboración propia

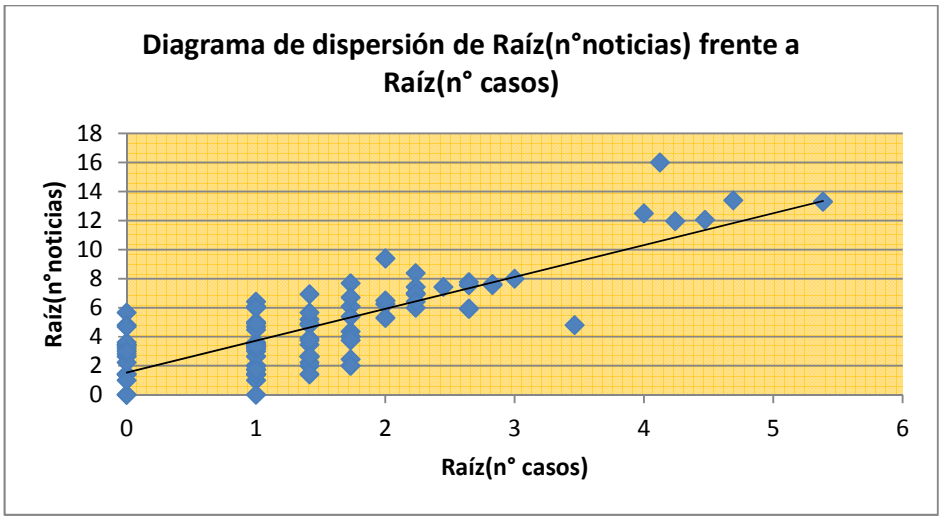

Figura 65: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre homicidios
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una relación estadísticamente significativa entre la raíz del número de casos y la raíz del número de noticias (con un p-valor aproximadamente a cero), la cual explicaría el 65,6% de la variabilidad en la raíz del número de noticias.

Los datos referentes a la región metropolitana muestran valores relativamente más altos que en el resto de las regiones, por lo que podría tener una influencia demasiado importante en los resultados. El estudio sin considerar los datos aportados por la región metropolitana muestra un coeficiente de correlación de 0,631 entre las variables analizadas, por lo que se decide mantener el modelo inicial, pues se conserva relativamente fuerte la probable relación lineal entre las variables. En la Figura 66 se observa que la ausencia de los valores relacionados a la región metropolitana no modifica la ecuación de la regresión de manera sustancial.

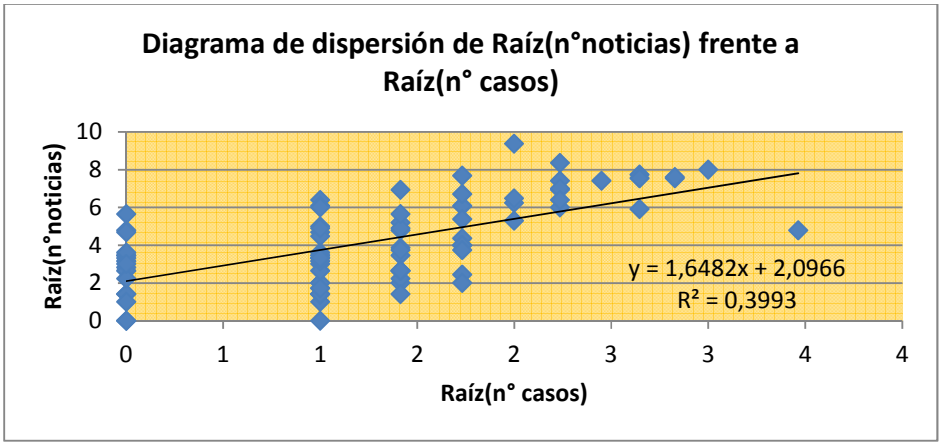

Figura 66: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre homicidios sin los datos de la XIII región
Fuente: Elaboración propia

4.6.5. Temática tránsito

La Figura 67 muestra, para la temática tránsito, una aparente relación lineal entre la raíz cuadrada del número de casos y la raíz cuadrada del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

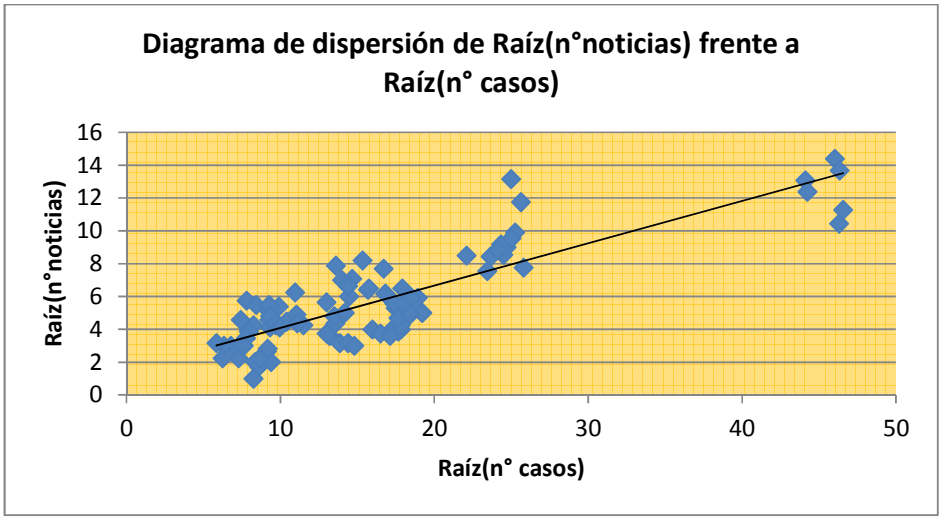

Figura 67: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre tránsito
Fuente: Elaboración propia

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$\sqrt{(N° \, Noticias)} = 1{,}513 + 0{,}258 * \sqrt{(N° \, Casos)}$$

R	R^2	Error Típico de la Estimación
0,855	0,731	1,509

Cuadro 41: Resumen del modelo de regresión lineal para la temática tránsito
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una relación estadísticamente significativa entre la raíz del número de casos y la raíz del número de noticias (con un p-valor aproximadamente a cero), la cual explicaría el 73,1% de la variabilidad en la raíz del número de noticias.

Los datos referentes a la región metropolitana muestran valores relativamente más altos que en el resto de las regiones, por lo que podría tener una influencia demasiado importante en los resultados. El estudio sin considerar los datos aportados por la región metropolitana muestra un coeficiente de correlación de 0,767 entre las variables analizadas, por lo que se decide mantener el modelo inicial, pues se conserva relativamente fuerte la probable relación lineal entre las variables. En la Figura 68 se observa que la ausencia de los valores relacionados a la región metropolitana no modifica la ecuación de la regresión de manera sustancial.

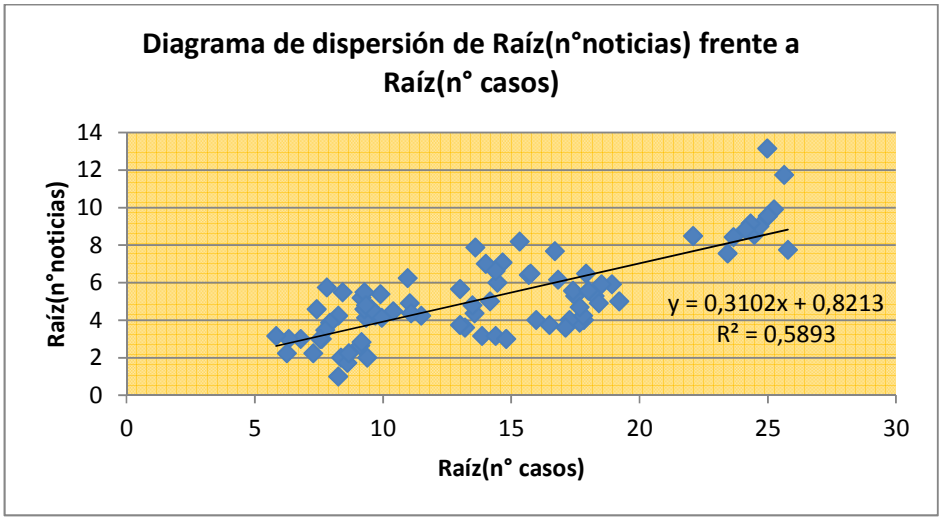

Figura 68: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre tránsito sin los datos de la XIII región
Fuente: Elaboración propia

4.6.6. Temática disturbios

La Figura 69 muestra, para la temática disturbios, una aparente relación lineal entre el logaritmo natural del número de casos y el logaritmo natural del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

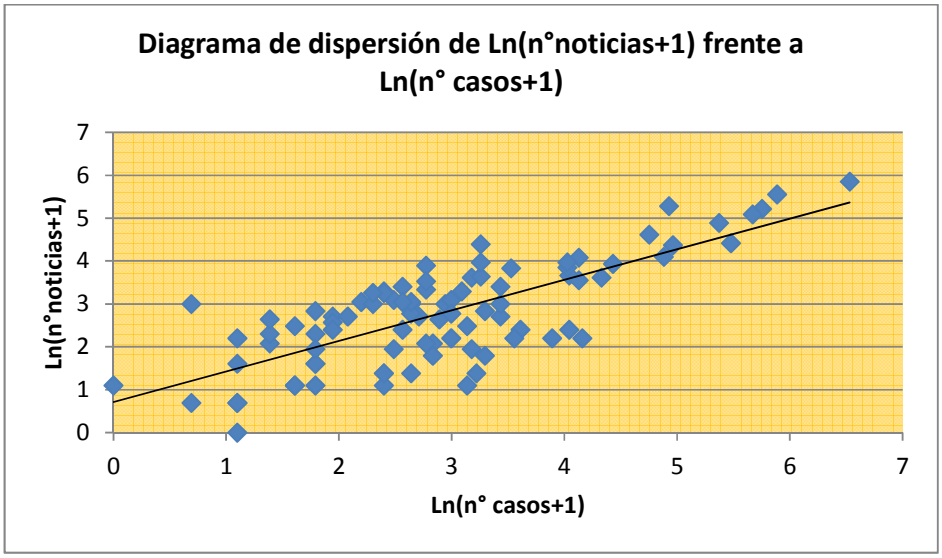

Figura 69: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre disturbios
Fuente: Elaboración propia

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$Ln(N°\ Noticias + 1) = \beta_0 + \beta_1 * Ln(N°\ Casos + 1) + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$Ln(N°\ Noticias + 1) = 0{,}710 + 0{,}713 * Ln(N°\ Casos + 1)$$

R	R^2	Error Típico de la Estimación
0,765	0,586	0,770

Cuadro 42: Resumen del modelo de regresión lineal para la temática disturbios
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una

relación estadísticamente significativa entre el logaritmo del número de casos y el logaritmo del número de noticias (con un p-valor aproximadamente a cero), la cual explicaría el 58,6% de la variabilidad en la raíz del número de noticias.

4.6.7. Temática incendios

La Figura 70 muestra, para la temática incendios, una aparente relación lineal entre logaritmo natural del número de casos y la raíz cuadrada del número de noticias, ambos mensuales. Se observa que, en general, a mayor número de casos, mayor es el número de noticias.

El modelo de regresión lineal simple estudiado es de la siguiente forma:

$$\sqrt{(N°\ Noticias)} = \beta_0 + \beta_1 * Ln(N°\ Casos + 1) + \varepsilon$$

, cuya estimación por el método MCO permite obtener la siguiente recta de regresión estimada:

$$\sqrt{(N°\ Noticias)} = -3{,}133 + 1{,}955 * Ln(N°\ Casos + 1)$$

R	R^2	Error Típico de la Estimación
0,750	0,562	1,587

Cuadro 43: Resumen del modelo de regresión lineal para la temática incendios
Fuente: Elaboración propia

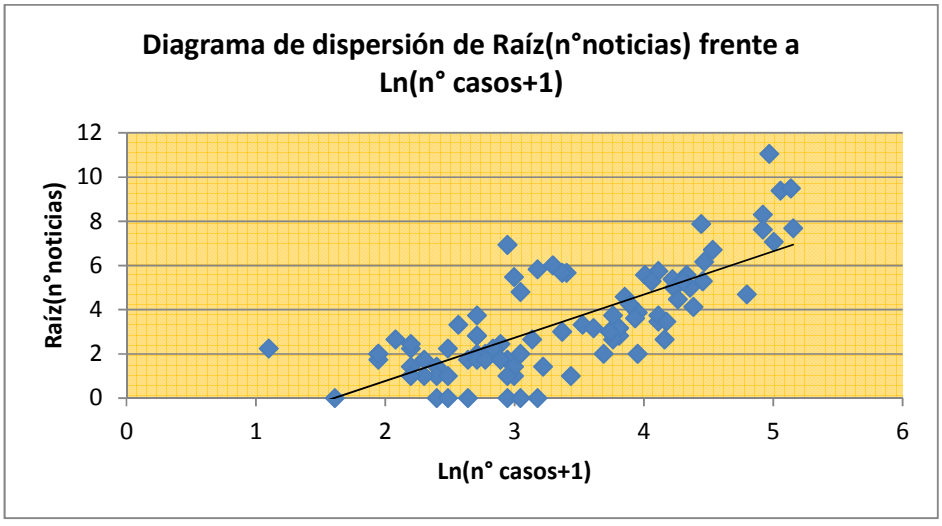

Figura 70: Diagrama de dispersión entre el n° de noticias y el n° de casos sobre incendios
Fuente: Elaboración propia

Este modelo cumple con los supuestos de normalidad, homocedasticidad e independencia de los errores. Del análisis de regresión se obtiene que existe una relación estadísticamente significativa entre el logaritmo del número de casos y la raíz del número de noticias (con un p-valor aproximadamente a cero), la cual explicaría el 56,2% de la variabilidad en la raíz del número de noticias.

4.7. Prototipo de herramienta para visualización geográfica

Se desarrolla el prototipo de una herramienta que permite visualizar de forma eficiente los niveles de casos y noticias por habitantes en cada una de las regiones del país, con el objetivo de verificar si la cobertura de las noticias para las distintas temáticas policiales guarda alguna relación con el volumen de casos que no ha sido percibida con las herramientas utilizadas previamente.

El prototipo se basa en las funcionalidades disponibles a partir de la utilización del API de JavaScript de Google Maps. Se delimitan las zonas geográficas de cada región del país y se designa un color de fondo para una región de acuerdo al nivel alcanzado por las variables que se analizan. Los datos son recopilados desde una hoja de cálculo (Excel) y están definidos por región, mes y temática policial.

Como medida de clasificación de los datos en estudio se utiliza su desviación estándar a la media, separando los datos en tres grupos: valores bajo 0,5 desviaciones estándar de la media, valores sobre 0,5 desviaciones estándar de la media y el resto de los valores (entre -0,5 y 0,5 desviaciones estándar de la media). Para el cálculo de la media y desviación estándar se utilizan los datos disponibles para cada variable por región y por mes (un total de 90 datos por temática a estudiar).Se realizan pruebas del prototipo para las temáticas drogas y robos, para dos meses distintos cada una.

En la Figura 71 y Figura 72, sobre la temática drogas, se puede visualizar que se forman grupos de regiones con distintos niveles de casos por habitantes a medida que se avanza en el territorio: en una zona norte el grupo con mayores niveles de casos, en una zona más central niveles medios de casos y en una zona sur niveles más bajos de casos. En general se constata una buena correspondencia entre los niveles de casos y los niveles de noticias para las distintas regiones.

En la Figura 73 y Figura 74, sobre la temática robos, se observa que tanto la distribución de los niveles de casos por habitantes como la distribución de los niveles de noticias por habitantes se mantienen relativamente constantes entre las regiones en los dos meses estudiados. Por otro lado, no se constata una buena correspondencia entre los niveles de casos y los niveles de noticias: regiones con niveles medios y bajos de casos están asociadas a niveles altos de noticias.

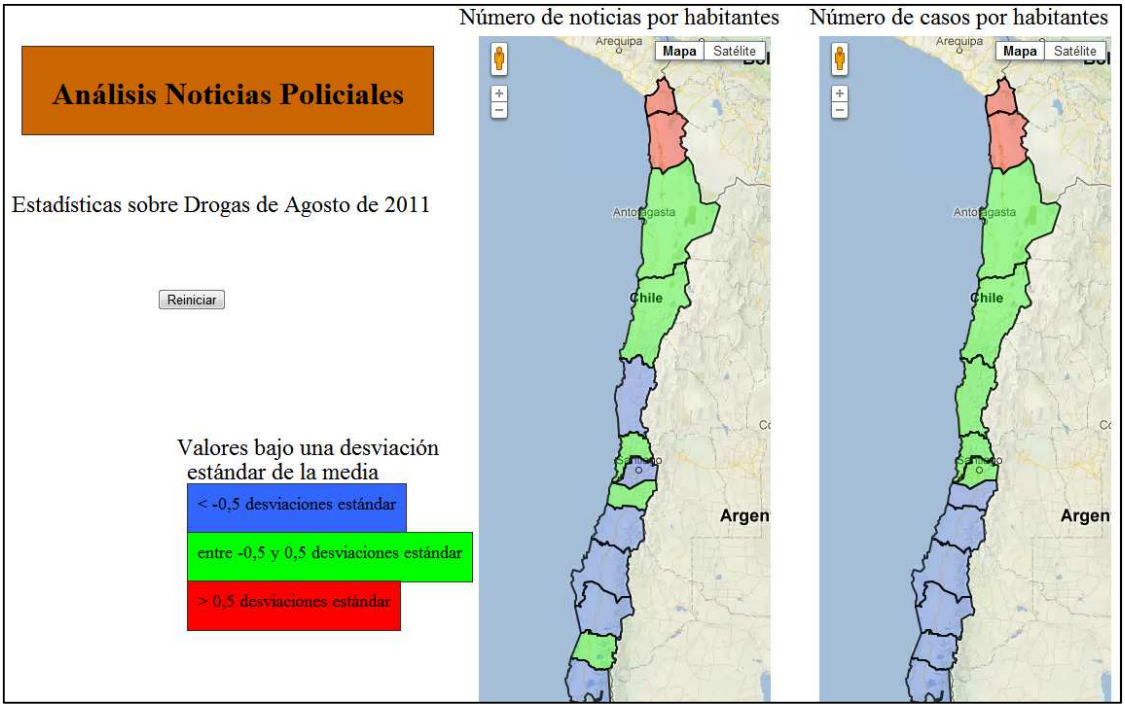

Figura 71: Visualización de estadísticas sobre drogas de agosto 2011
Fuente: Elaboración propia

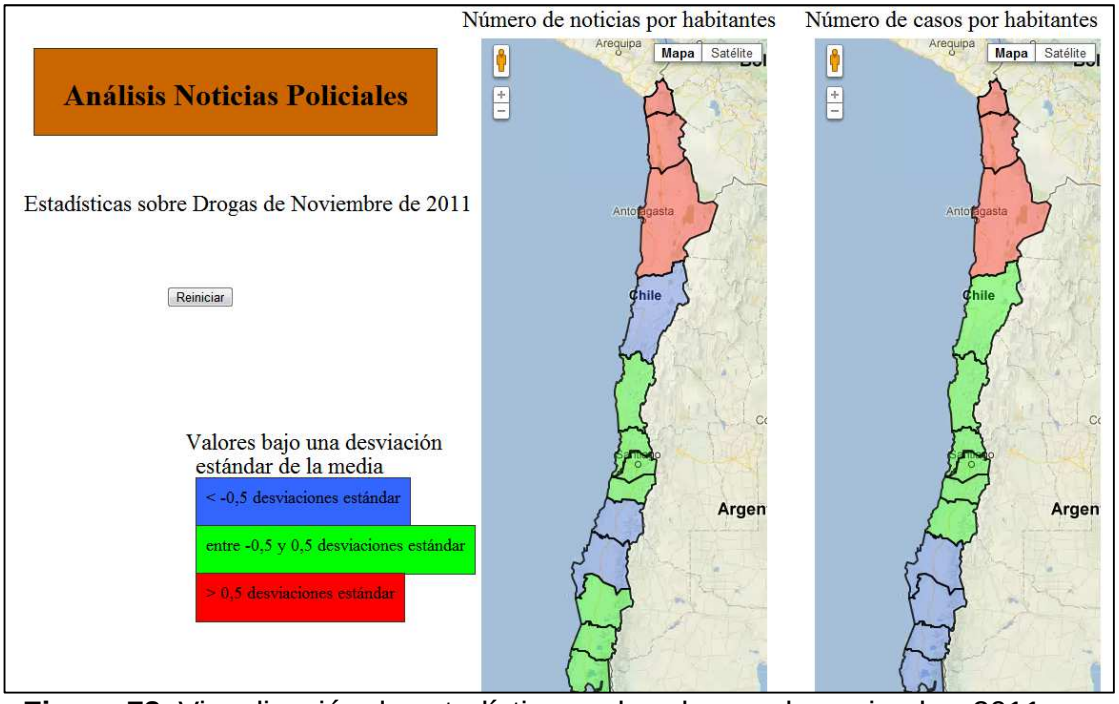

Figura 72: Visualización de estadísticas sobre drogas de noviembre 2011
Fuente: Elaboración propia

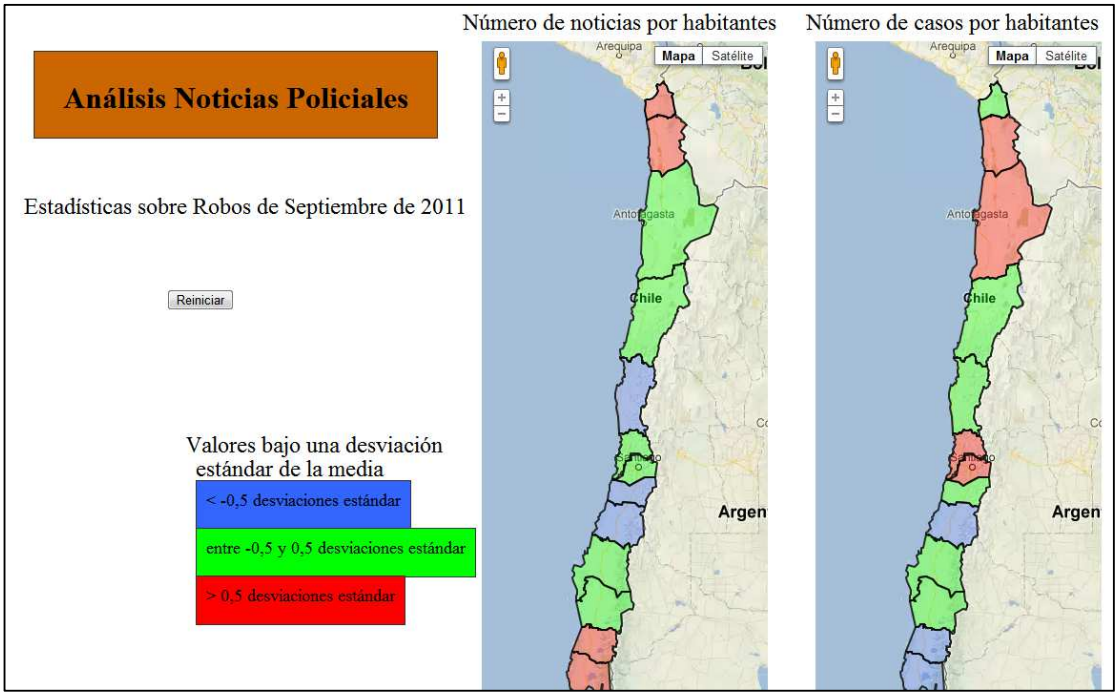

Figura 73: Visualización de estadísticas sobre robos de septiembre 2011
Fuente: Elaboración propia

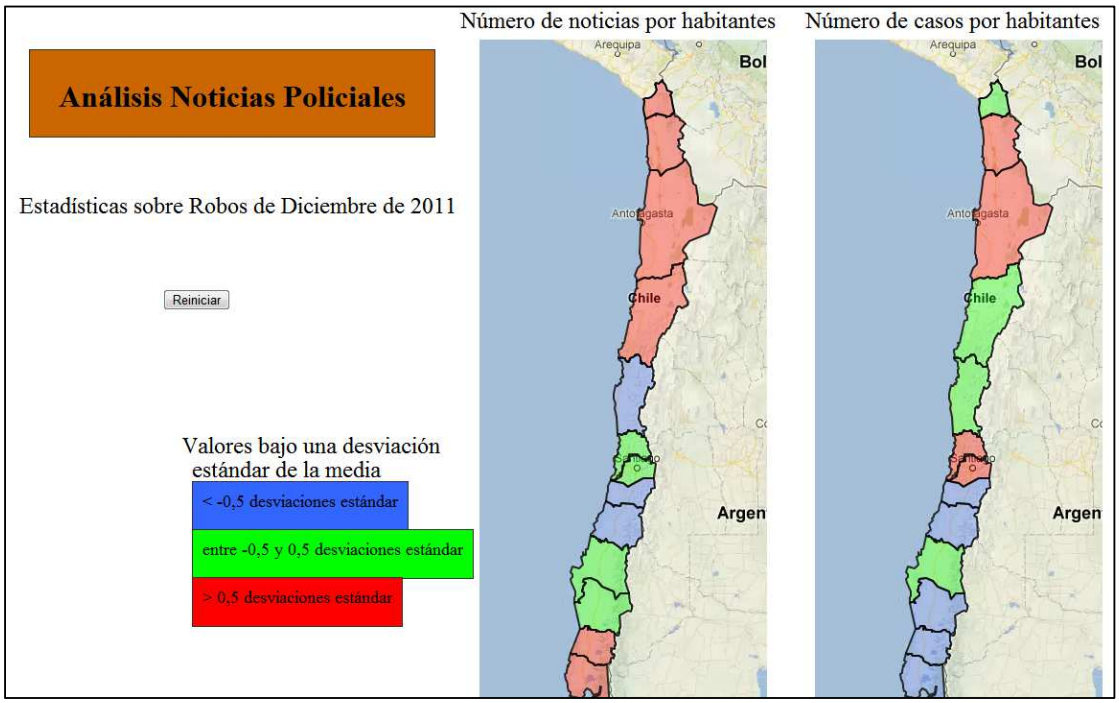

Figura 74: Visualización de estadísticas sobre robos de diciembre 2011
Fuente: Elaboración propia

5. Conclusiones

En este capítulo se realiza un balance del trabajo de título realizado, el cual está basado principalmente en la evaluación del grado de cumplimiento de los objetivos definidos inicialmente y un balance de los resultados obtenidos. También se presentan las recomendaciones para trabajos futuros.

5.1. Visión general del estudio

Las noticias policiales han sido continuamente objeto de análisis, tanto en estudios nacionales como internacionales, debido al efecto (y probable distorsión) que tendrían sobre los niveles de preocupación frente al delito por parte de la población. La minería de textos es un tema que cada vez toma mayor protagonismo, existiendo eficientes herramientas para el manejo de documentos de textos. Los artículos noticiosos poseen valiosa información que muchas veces no es explotada, además pueden complementar y enriquecer diversos estudios sobre medición de percepción en diferentes temas. El análisis de noticias es un proceso manual, intensivo en tiempo y recursos, que en general se realiza sin aprovechar las numerosas herramientas tecnológicas que se disponen actualmente para cumplir con esta tarea de forma eficiente, disminuyendo notoriamente los potenciales errores humanos y permitiendo trabajar con crecientes volúmenes de datos. La aplicación de técnicas de minería de textos sobre un conjunto de noticias policiales permitiría descubrir interesantes patrones a partir de los textos

El principal objetivo de esta memoria es el diseño de una metodología apoyada en técnicas de minería de textos que permita desarrollar un análisis eficiente de noticias policiales, utilizando la nueva e interesante información descubierta para posteriores estudios complementados con información externa. La metodología de investigación está basada en las etapas del modelo de procesos CRISP-DM, el que permite diseñar, de forma esquematizada, las distintas tareas involucradas para cumplir con los objetivos planteados, para lo cual se debe integrar diferentes pasos como la comprensión de la naturaleza del problema, la comprensión y preparación de los datos, la construcción y evaluación de los modelos y los posteriores desarrollos a partir del conocimiento generado.

Uno de los principales problemas enfrentados en el trabajo con documentos de textos es su alta dimensionalidad al representar cada texto como una bolsa de palabras, por lo que se asignan grandes esfuerzos a formar una lista de stop words, las bases del proceso stemming y la selección supervisada y no supervisada de atributos. Una de las tareas fundamentales de la investigación es la correcta identificación de temáticas policiales a partir del contenido de las noticias utilizando modelos de

clustering, aunque primero se debe detectar si una noticias es o no de carácter policial. Los modelos de clasificación son evaluados usando criterios como accuracy, precisión, recall y F-measure, en tanto los modelos de clustering son evaluados usando el índice Davies-Bouldin y mediante inspección del contenido.

Uno de puntos que facilita esta investigación es la disponibilidad de la información a utilizar, dado que las noticias se obtienen de forma relativamente sencilla haciendo uso de métodos simples de extracción de contenido favorecido por el formato RSS y, por otro lado, la información externa complementaria, como las estadísticas de casos policiales, son de acceso público.

La metodología implementada en este proyecto permite cumplir exitosamente con los objetivos propuestos inicialmente, describiendo, de forma estructurada, una secuencia de etapas que deben ser superadas. De esta forma, el modelo de procesos, muchas veces complejo y que requiere de múltiples herramientas, facilita el entendimiento y manejo de las distintas interacciones entre las etapas involucradas. La metodología entrega un consejo para cada tarea, sirviendo, adicionalmente, como una lista para la verificación del cumplimiento de ellas, evitando que ningún punto relevante sea olvidado.

Una de las principales ventajas de la metodología propuesta es que no requiere necesariamente de personal altamente especializado para su aplicación. Así mismo, dado que las tareas que involucran cada etapa están previamente definidas, se pueden realizar mejoras parciales para las secciones que lo requieran, sin necesidad de afectar el resto del proceso.

5.2. Balance de los resultados obtenidos

A partir de la aplicación de la metodología propuesta en esta investigación a un conjunto de noticias, obtenidas de cuatro medios de prensa nacionales recopiladas durante seis meses, se obtienen el siguiente balance general:

- La clasificación de noticias entre policiales y no policiales obtiene niveles de desempeño bastante aceptables: accuracy de 93,73% y F-measure de 92,10%.

- El resultado del proceso de identificación de temáticas policiales presentes en las noticias se considera bastante aceptable, dado que se detectan para los cuatro medios de prensa en estudio las mismas temáticas policiales. Las temáticas son fácilmente identificables y distinguibles entre sí al examinar las palabras más relevantes que componen cada grupo.

- La caracterización de una temática utilizando herramientas como reglas de asociación y tag clouds resultan bastante útiles y complementarias para distinguir y validar su contenido.

- Se mantiene una relativa proporcionalidad entre la cantidad de noticias policiales y el número de habitantes para la mayoría de las regiones. En regiones como IV y VI se detectan niveles de noticias por habitantes muy bajos en comparación al resto, mientras que la XV presenta el nivel más alto.

- La cobertura de las distintas temáticas policiales presenta variaciones importantes entre cada una de las regiones. Entre las temáticas policiales que presentan mayor cobertura se encuentran las noticias sobre robos y tránsito, mientras que la menor cobertura es para las noticias sobre incendios y delitos sexuales.

- En temáticas como delitos sexuales, homicidios y tránsito se observa que el nivel de casos por habitantes es relativamente constante entre las distintas regiones, mientras que los niveles de noticias policiales por habitantes sobre estas mismas temáticas presentan una mayor variabilidad entre las regiones.

- A partir del análisis de regresiones lineales simples se concluye que el número de casos podría explicar una proporción importante de la variabilidad del número de noticias para las distintas temáticas policiales. Es probable que para obtener resultados que describan mejor el comportamiento del nivel de noticias policiales deban agregarse otros factores no considerados en el estudio.

- Se detecta que para las temáticas homicidios e incendios existe la mayor cantidad de noticias informadas por cada caso real registrado, mientras que la menor cantidad de noticias por caso real se detecta para las temáticas robos y drogas.

- El prototipo de la herramienta para visualización de datos georreferenciados cumple con las expectativas planteadas inicialmente, aunque puede aprovecharse más aplicando mejoras. La visualización permite revisar distintos datos simultáneamente de forma eficiente, permitiendo detectar nuevos patrones que con otras herramientas fueron ignorados.

5.3. Limitaciones de los resultados

Este trabajo ha dado un importante paso al mostrar la factibilidad de la identificación de temáticas policiales dentro de las noticias, evaluando técnicas relativamente clásicas de aprendizaje supervisado y no supervisado, tales como K-nn, Naive Bayes y K-means. Con técnicas más avanzadas se podrían alcanzar posiblemente mejores resultados.

Por otro lado, el proceso de reducción del número de atributos es uno de los procesos que mayor sensibilidad muestra sobre los resultados, por lo que gran influencia tienen los procesos de selección de atributos, sobre el cual se realiza un número limitado de evaluaciones con distintas configuraciones.

Otra de las limitaciones sobre los resultados es el conjunto de noticias tomado como muestra para aplicar la metodología en estudio, dado que la recopilación de datos se realiza durante seis meses, los resultados pueden estar acotados a cierta estacionalidad y no reflejar el comportamiento de las noticias en el resto del año (por ejemplo no contemplar ciertas temáticas policiales que puedan tener cobertura durante el período no evaluado). Otro de los puntos conflictivos es la alta cobertura de una temática debido al acontecer noticioso de un par de meses, pero que, en general, puede no tener presencia destacada en los medios (ejemplo de esto son las noticias sobre disturbios relacionadas probablemente con las actividades de un movimiento estudiantil durante el período de estudio).

El desarrollo de la herramienta de visualización, propuesta en esta memoria, permite una excelente comprensión del comportamiento espacial en relación a las noticias policiales, el que podría beneficiarse con un conocimiento más acabado en lenguajes de programación (php y html).

5.4. Recomendaciones para trabajos futuros

La metodología presentada en este trabajo puede ser planteada para el monitoreo continuo de las noticias policiales, lo que permitiría realizar análisis sistemáticos en el tiempo, dado que las distintas tareas involucradas están plenamente definidas y ya aplicadas a un conjunto de noticias de prueba con buenos resultados. Haciendo uso de esta metodología, que incorpora el uso de herramientas de minería de textos, se logra un manejo eficiente de datos no estructurados, además de tener la capacidad de procesar crecientes niveles de información.

Otro de los puntos sobre los cuales se podría trabajar es la definición del análisis basado en la información para las provincias y/o comunas del país, adicionalmente a la configuración regional aplicada en este trabajo, lo que permitiría conocer el comportamiento detallado de las noticias policiales comparado con los casos reales registrados.

Finalmente se plantea la posibilidad de implementar una herramienta capaz de integrar las distintas etapas de la metodología propuesta (o la mayoría de ellas), que permita manejar grandes volúmenes de datos, definir las principales configuraciones de las tareas involucradas en el proceso y presentar los resultados de forma sencilla al analista.

6. Bibliografía

[1] G. ADAM, C. BOURAS and V. POULOPOULOS. Efficient extraction of news articles based on RSS crawling. In International Conference on Machine and Web Intelligence (ICMWI), pages 1-7, October 2010.

[2] C. C. AGGARWAL and C. ZHAI. Mining text data. Springer, 2012.

[3] R. AGRAWAL, T. IMIELINSKI and A. SWAMI. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 22(2): 207-216, June 1993.

[4] E. AKPINAR and N. USUL. Geographic information systems technologies in crime analysis and crime mapping. In P. Saxena and Ravi Kumar Jain, editors, Geographic Information System: an introduction. Icfai University Press, 2007.

[5] C. ALLENDE. El peso del temor a la delincuencia y sus factores determinantes en la población urbana chilena. En Primer Simposio Nacional de Investigación sobre Violencia y Delincuencia, publicado por ISUC/FPC, páginas 199-220, 2004.

[6] S. ANANYAN and V. KOLLEPARA. Crime pattern analysis, Megaputer case study in text mining. Megaputer Intelligence Inc., 2002.

[7] A. AZEVEDO and M. F. SANTOS. KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining, pages 182-185, 2008.

[8] A. M. BEHAR y P. LUCILLI. Mapa del delito de la Ciudad Autónoma de Buenos Aires. Terceras Jornadas de Jóvenes Investigadores, Instituto Gino Germani, 2003.

[9] B. BERENDT. Text mining for news and blogs analysis. In C. Sammut & G. Webb (Eds.), Encyclopedia of Machine Learning, pages 968-972. Springer, 2011.

[10] A. BERGO. Text categorization and prototypes. Technical report, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2001.

[11] M. W. BERRY and J. KOGAN. Text mining applications and theory. John Wiley & Sons, United Kingdom, 2010.

[12] M. BRAMER. Principles of data mining (undergraduate topics in computer science). Springer, London, UK, 2007.

[13] M. BROWNE y V. TOMICIC. Crimen y temor: el rol de los medios. En Cuadernos de Información Pontificia Universidad Católica de Chile, n° 20, páginas 21-36, 2007.

[14] M. BRUN, C. SIMA, J. HUA, J. LOWEY, B. CARROLL, E. SUH and E. R. DOUGHERTY. Model-based evaluation of clustering validation measures. In Pattern Recognition, 40(3): 807–824, March 2007.

[15] P.R. CANTER. Geographic information systems and crime analysis in Baltimore Country, Maryland. In D. Weisburd and McEwen T., editors, Crime Mapping and Crime Prevention, pages 157-190. Criminal Justice Press, Monsey, New York, USA, 1998.

[16] P. CHAPMAN, J. CLINTON, R. KERBER, T. KHABAZA, T. REINARTZ, C. SHEARER and R. WIRTH. CRISP-DM 1.0 Step-by-step data mining guide. Technical Report, CRISP-DM Consortium, August 2000.

[17] M. CHAU, J. J. XU and H. CHEN. Extracting meaningful entities from police narrative reports. In Proceedings of the 2002 Annual National Conference on Digital government research, pages 1-5, May 2002.

[18] H. CHEN, W. CHUNG, J. XU, G. WANG, Y. QIN and M. CHAU. Crime data mining: a general framework and some examples. In IEEE Computer, 37(4): 50-56, April 2004.

[19] H. CHERFI, A. NAPOLI and Y. TOUSSAINT. Towards a text mining methodology using frequent itemsets and association rule extraction. In Proc. Fourth Int'l Conf. Journées de l'Informatique Messine (JIM '03) on Knowledge Discovery and Discrete Math., pages 285-294, 2003.

[20] M. DASH and H. LIU. Feature selection for classification. In International Journal of Intelligent Data Analysis, 1(3):131-156, 1997.

[21] C. DASTRES, C. SPENCER, E. MUZZOPAPPA y C. SÁEZ. La construcción de noticias sobre seguridad ciudadana en prensa escrita y televisión. ¿Posicionamiento, distorsión o comprensión?. En Colección Seguridad Ciudadana y Democracia, n° 2, Instituto de Asuntos Públicos, Universidad de Chile, 2005.

[22] D. L. DAVIES and D. W. BOULDIN. In cluster separation measure. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2): 95-104, 1979.

[23] I. S. DHILLON and D. S. MODHA. Concept decompositions for large sparse text data using clustering. In Machine Learning, 42(1-2): 143-175, 2001.

[24] I. S. DHILLON, J. KOGAN and M. NICHOLAS. Feature selection and document clustering. In M. Berry, editor, A Comprehensive Survey of Text Mining, pages 73-100. Springer, New York, 2004.

[25] E. ESPINAR Y R. RUIZ. El crimen en los programas informativos de la TV española. En Cuadernos de Información Pontificia Universidad Católica de Chile, nº 26, páginas 65-76, 2010.

[26] P. FALINOUSS. Stock trend prediction using news articles: a text mining approach. Master's thesis, Luleå University of Technology, 2007.

[27] U. M. FAYYAD, G. PIATETSKY-SHAPIRO and P. SMYTH. From data mining to knowledge discovery in databases. In AI Magazine, 17(3): 37- 54, 1996.

[28] U. M. FAYYAD, G. PIATETSKY-SHAPIRO and P. SMYTH. The KDD process for extracting useful knowledge from volumes of data. In Communications of the ACM: 39(11): 27-34, 1996.

[29] R. FELDMAN and J. SANGER. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, New York, USA, 2007.

[30] G. FORMAN. An extensive empirical study of feature selection metrics for text classification. In Journal of Machine Learning Research, vol. 3, pages 1289-1305, 2003.

[31] M. GAETE. La seguridad ciudadana en los noticieros de televisión. En Informe Político N°80, Libertad y Desarrollo, 2003.

[32] G. GERBNER and L. GROSS. Living with television: the violence profile. In Journal of Communication, 26(2), 173-199, 1976.

[33] G. GERBNER. Cultivation analysis: an overview. In Mass Communication Research 3-4, 175-194, 1998.

[34] D. N. GUJARATI and D C. PORTER. Econometría. McGRAW-HILL, México, quinta edición, 2010.

[35] V. GUPTA and G. S. LEHAL. A survey of text mining techniques and applications. In Journal of Emerging Technologies in Web Intelligence, 1(1): 60-76, August 2009.

[36] I. GUYON and A. ELISSEEFF. An introduction to variable and feature selection. In The Journal of Machine Learning Research, vol. 3, pages 1157-1182, January 2003.

[37] B. HAGHIGHI and J. SORENSEN. America's fear of crime. In Flanagan & Longmire, editors, Americans View Crime and Justice: A National Public Opinion Survey, pages 16-30. Sage Publications, USA, 1996.

[38] M. HALKIDI, Y. BATISTAKIS and M. VAZIRGIANNIS. On clustering validation techniques. In Journal of Intelligent Information Systems, 17(2-3): 107-145, 2001.

[39] J. HAN and M. KAMBER. Data mining: concepts and techniques. Morgan Kaufmann, second edition, 2006.

[40] Y. HASSAN-MONTERO and V. HERRERO-SOLANA. Improving tag-clouds as visual information retrieval interfaces. In InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies, 2006.

[41] C. HAYES and P. AVESANI. Using tags and clustering to identify topic-relevant blogs. In Proceedings of International Conference on Weblog and Social Media (ICWSM-07), 2007.

[42] H. R. HOLZMAN, W. D. WHEATON and D. P. CHREST. Partnering with the police to prevent crime using geographic information systems: a guide for housing authorities and other community stakeholders. October 2003.

[43] A. HOTHO, A. NÜRNBERGER and G. PAAß. A brief survey of text mining. In LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20(1):19-62, May 2005.

[44] A. HUANG. Similarity measures for text document clustering. In The Proceedings of New Zealand Computer Science Research Student Conference, 2008.

[45] A.K. JAIN, M.N. MURTY and P.J. FLYN. Data clustering: a review. In ACM Computing Surveys, 31(3):264–323, 1999.

[46] P. JAJOO. Document clustering. Master's thesis, Indian Institute of Technology Kharagpur, 2008.

[47] M. KLEMETTINEN, H. MANNILA, P. RONKAINEN, H. TOIVONEN and A.I. VERKAMO. Finding interesting rules from large sets of discovered association rules. In Proceedings of the Third International Conference on Information and Knowledge Management (CIKM 1994), pages 401–407, 1994. ACM Press.

[48] J. KOGAN, C. NICHOLAS and V. VOLKOVICH. Text mining with information-theoretic clustering. In IEEE Computing in Science and Engineering, 5(6):52-59, 2003.

[49] C. LEGÁNY, S. JUHÁSZ and A. BABOS. Cluster validity measurement techniques. In Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'06), pages 388-393, February 2006.

[50] L. LIU, J. KANG, J. YU and Z. WANG. A comparative study on unsupervised feature selection methods for text clustering. In Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, pages 597–601, 2005.

[51] T. LIU, S. LIU and Z. CHEN. An evaluation on feature selection for text clustering. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), 2003.

[52] L. LLOYD, D. KECHAGIAS and S. SKIENA. Lydia: a system for large-scale news analysis. In String Processing and Information Retrieval: 12th International Conference, SPIRE 2005, Buenos Aires, Argentina. Lecture Notes in Computer Science, vol. 3772, pages 161–166, November 2005.

[53] U. MAULIK and S. BANDYOPADHYAY. Performance evaluation of some clustering algorithms and validity indices. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pages 1650 -1654, December 2002.

[54] M. MCCOMBS and D.L. SHAW. The agenda-setting function of mass media. In Political Opinion Quarterly, 36(2): 176-187, 1972.

[55] M. MCCOMBS. Estableciendo la agenda: el impacto de los medios en la opinión pública y en el conocimiento. Editorial Paidós, Barcelona, España, 2006.

[56] A. MEHLER, Y. BAO, XIN LI, YUE WANG and S. SKIENA. Spatial analysis of news sources. In IEEE Transactions on Visualization and Computer Graphics, 12(5):765-772, September 2006.

[57] G.W. MILLIGAN and M.C. COOPER. An examination of procedures for determining the number of clusters in a data set. In Psychometrika, 50(2): 159-179, 1985.

[58] M. MONTES Y GÓMEZ, A. GELBUKH and A. LÓPEZ. Mining the news: trends, associations and deviations. En Computación y Sistemas, Revista Iberoamericana de Computación, 5(1): 14-24, 2001.

[59] K. MUGUNTHADEVI, S. C. PUNITHA and M. PUNITHAVALLI. Survey on feature selection in document clustering. In International Journal on Computer Science and Engineering (IJCSE), 3(3): 1240-1244, March 2011.

[60] S. V. NATH. Crime pattern detection using data mining. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pages 41-44, 2006.

[61] A. NIKFARJAM, E. EMADZADEH AND S. MUTHAIYAH. Text mining approaches for stock market prediction. In The 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 4, pages 256–260, New York, USA, February 2010.

[62] K. NORVAG and R. OYRI. News item extraction for text mining in web newspapers. In Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration, pages 195-204, 2005.

[63] F. OKWANGALE and P. OGAO. Survey of data mining methods for crime analysis and visualization. In Special Topics in Computing and ICT Research: Advances in Systems Modelling and ICT Applications, vol. 2, pages 221-226, 2006.

[64] A. ÖZGÜR, L. ÖZGÜR and T. GÜNGÖR. Text categorization with class-based and corpus-based keyword selection. In Proceedings of the 20th International Symposium on Computer and Information Sciences, volume 3733 of Lecture Notes in Computer Science, pages 606-615, October 2005.

[65] A. PÉREZ GARCÍA-PLAZA, A. ZUBIAGA, V. FRESNO and R. MARTÍNEZ. Reorganizing clouds: a study on tag clustering and evaluation. In Expert Systems with Applications, 39(10): 9483-9493, August 2012.

[66] M. PLASSE, N. NIANG, G. SAPORTA, A. VILLEMINOT and L. LEBLOND. Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. In Computational Statistics and Data Analysis, 52(1): 596-613, 2007.

[67] H. QIAO and B. EDWARDS. A data clustering tool with cluster validity indices. In International Conference on Computing, Engineering and Information, pages 303-309, April 2009.

[68] J. O. RAWLINGS, S. G. PANTULA and D. A. DICKEY. Applied regression analysis, a research tool. Springer-Verlag, New York, second edition, 1998.

[69] E. RENDÓN, I. ABUNDEZ, A. ARIZMENDI and E. M. QUIROZ. Internal versus external cluster validation indexes. In International Journal of Computers and Communications. 5(1): 27–34, 2011.

[70] C. ROBARDET, B. CRÉMILLEUX and J.-F. BOULICAUT. Characterization of unsupervised clusters by means of the simplest association rules: an application for child's meningitis. In F. Lyon (ed.), Proceedings IDAMAP 2002 co-located with ECAI 2002, pages 61–66, 2002.

[71] G. SALTON, A. WONG and C. S. YANG. A vector space model for automatic indexing. In Communications of the ACM, 18(11): 613-620, Nov. 1975.

[72] H. SAYYADI, S. SALEHI and H. ABOLHASSANI. Survey on news mining tasks. In Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, pages 219-224. Springer, 2007.

[73] F. SEBASTIANI. 2002. Machine learning in automated text categorization. In ACM Computing Surveys, 34 (1): 1-47, March 2002.

[74] S. SOTO NAVARRO. La influencia de los medios en la percepción social de la delincuencia. En Revista Electrónica de Ciencia Penal y Criminología, n°7, 2005.

[75] M. STEINBACH, G. KARYPIS and V. KUMAR. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[76] A. STREHL, J. GHOSH and R. J. MOONEY. Impact of similarity measures on web-page clustering. In Proc. AAAI Workshop on AI for Web Search (AAAI 2000), pages 58-64. AAAI/MIT Press, July 2000.

[77] B. TANG, M. SHEPHERD, E. MILIOS and M. I. HEYWOOD. Comparing and combining dimension reduction techniques for efficient text clustering. In Proceedings of SIAM International Workshop on Feature Selection for Data Mining, pages 17-26, 2005.

[78] T. TYLER and F. COOK. The mass media and judgements of risk: distinguishing impact on personal and societal level judgements. In Journal of Personality and Social Psychology, vol. 47, pages 693-708, 1984.

[79] F. VALENGA, E. FERNÁNDEZ, H. MERLINO, D. RODRÍGUEZ, C. PROCOPIO, P. BRITOS y R. GARCÍA-MARTÍNEZ. Minería de datos aplicada a la detección de patrones delictivos en argentina. Proceedings VII Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento, páginas 31-39, 2008.

[80] R.C.P. VAN DER VEER, H.T. ROOS and A. VAN DER ZANDEN. Data mining for intelligence led policing. From Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 2009.

[81] S. RÍOS, J. VELÁSQUEZ, H. YASUDA and T. AOKI. Conceptual classification to improve a web site content. In Proceedings of the 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'06), pages 869-877.

[82] VIDHYA K. A. and G. AGHILA. Text mining process, techniques and tools: an overview. In International Journal of Information Technology and Knowledge Management 2(2): 613-622, 2010.

[83] M. VIJAYA KUMAR and C. CHANDRASEKAR. GIS technologies in crime analysis and crime mapping. In International Journal of Soft Computing and Engineering (IJSCE), 1(5): 115-121, November 2011.

[84] S. M. WEISS, N. INDURKHYA and T. ZHANG. Fundamentals of predictive text mining. Springer, London, UK, 2010.

[85] Y. YANG and X. LIU. A re-examination of text categorization methods. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, pages 42-49, 1999.

[86] Y. YANG and J. O. PEDERSEN. A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 412-420, July 1997.

[87] K. R. ŽALIK and BORUT ŽALIK. Validity index for clusters of different sizes and densities. In Pattern Recognition Letters, 32(2): 221–234, January 2011.

[88] Q. ZHAO and S. BHOWMICK. Association rule mining: a survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003.

[89] S. ZHONG. Efficient online spherical k-means clustering. In Proceedings of IEEE International Joint Conference on Neural Networks, vol.5, pages 3180-3185, 2005.

7. Anexos

7.1. N° de noticias y n° de casos mensuales por región según temática policial

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	162	21	172	11	219	17
I	166	19	347	19	399	23
II	270	17	176	22	234	30
III	34	13	47	7	56	6
IV	56	3	65	2	52	7
V	213	40	206	63	243	38
VI	33	12	45	14	84	17
VII	49	11	50	10	35	17
VIII	70	23	84	29	82	28
IX	16	13	20	14	13	33
XIV	5	15	10	7	4	3
X	14	16	15	5	19	12
XI	3	1	5	4	9	1
XII	1	1	2	2	6	0
XIII	1818	96	1792	83	1495	81

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	164	17	113	19	165	19
I	353	38	415	28	500	26
II	268	45	271	36	279	29
III	49	7	34	4	38	7
IV	73	9	49	12	59	5
V	232	39	248	56	245	43
VI	126	12	187	18	148	16
VII	69	10	65	5	44	17
VIII	84	25	73	28	87	38
IX	29	28	36	23	24	23
XIV	15	1	13	7	13	13
X	19	4	17	14	13	19
XI	1	1	2	0	1	4
XII	3	1	3	2	3	4
XIII	2297	100	2219	130	1809	107

Cuadro 44: N° de noticias y n° de casos mensuales sobre drogas
Fuente: Elaboración propia

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	257	9	251	13	249	20
I	859	15	765	16	754	28
II	1188	31	1322	35	1183	20
III	446	12	484	9	463	12
IV	899	7	928	11	846	7
V	3343	73	3420	69	3248	55
VI	1350	10	1324	6	1281	11
VII	1163	19	1246	15	988	16
VIII	2976	46	3278	82	3044	62
IX	1072	35	1327	30	1147	38
XIV	385	20	467	27	392	33
X	1057	61	1219	50	958	56
XI	90	2	112	1	146	3
XII	81	4	72	2	83	1
XIII	13660	186	14467	185	13953	226

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	285	15	315	10	254	12
I	833	21	670	13	747	20
II	1246	27	1219	36	1156	41
III	531	17	473	9	430	17
IV	836	15	875	14	875	12
V	3119	72	3014	84	2951	90
VI	1121	21	978	29	834	15
VII	1005	21	953	26	819	25
VIII	2887	56	2591	74	2463	89
IX	1025	61	955	53	935	41
XIV	377	19	319	23	298	23
X	979	55	851	57	818	56
XI	100	6	89	0	63	2
XII	105	5	106	4	109	5
XIII	13637	260	13324	264	12298	260

Cuadro 45: N° de noticias y n° de casos mensuales sobre robos
Fuente: Elaboración propia

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	13	3	7	10	9	12
I	19	3	21	7	22	9
II	26	14	35	17	39	17
III	17	5	11	7	15	8
IV	25	2	34	7	29	5
V	92	21	93	38	95	17
VI	43	2	39	12	48	7
VII	55	22	50	10	52	10
VIII	91	22	89	25	105	29
IX	30	35	50	28	48	30
XIV	21	12	24	10	24	7
X	37	24	61	26	65	29
XI	3	3	3	4	4	2
XII	9	1	7	3	6	4
XIII	365	41	418	52	396	54

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	13	9	20	8	11	11
I	16	4	26	7	18	7
II	31	15	42	16	35	16
III	20	8	19	6	12	4
IV	34	9	36	7	17	6
V	103	24	99	42	103	37
VI	73	2	47	12	60	5
VII	69	4	53	9	52	10
VIII	134	19	139	40	123	38
IX	62	16	70	31	71	35
XIV	19	1	25	5	26	8
X	69	12	59	27	62	19
XI	4	0	4	1	4	1
XII	14	0	7	1	9	0
XIII	451	72	545	83	474	86

Cuadro 46: N° de noticias y n° de casos mensuales sobre delitos sexuales
Fuente: Elaboración propia

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	1	13	0	12	0	13
I	1	7	1	11	2	7
II	1	24	2	27	0	23
III	1	2	1	11	0	7
IV	1	3	1	1	3	4
V	5	48	1	37	3	59
VI	3	14	3	16	12	23
VII	2	23	3	29	3	37
VIII	8	57	8	58	7	35
IX	4	88	1	41	1	25
XIV	0	22	1	9	1	22
X	3	45	4	42	2	48
XI	1	4	0	2	1	2
XII	0	1	0	5	1	0
XIII	16	156	20	145	18	143

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	1	12	2	15	0	9
I	1	3	2	5	0	10
II	4	28	2	14	1	20
III	2	7	0	8	0	11
IV	0	2	1	4	2	7
V	6	55	5	55	5	70
VI	2	12	5	36	3	6
VII	3	19	0	32	3	16
VIII	9	64	7	60	7	57
IX	2	24	4	39	1	36
XIV	1	11	1	10	0	12
X	5	49	5	41	2	32
XI	0	0	2	2	2	4
XII	1	12	1	2	0	0
XIII	29	177	17	256	22	179

Cuadro 47: N° de noticias y n° de casos mensuales sobre homicidios
Fuente: Elaboración propia

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	84	27	98	29	94	18
I	91	21	71	30	91	19
II	196	49	185	62	169	32
III	61	33	55	21	63	15
IV	174	13	169	14	184	19
V	665	60	609	81	624	173
VI	318	16	312	15	255	16
VII	306	28	369	25	321	42
VIII	599	73	592	84	560	71
IX	358	35	333	28	303	31
XIV	132	18	68	18	86	30
X	246	41	248	42	279	59
XI	53	5	46	9	40	9
XII	88	4	70	4	58	9
XIII	2143	109	2168	127	1946	171

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	122	24	120	39	99	17
I	83	7	86	21	108	20
II	209	36	182	23	201	25
III	60	12	87	23	123	19
IV	192	10	207	10	219	9
V	657	138	582	78	637	98
VI	298	16	272	14	293	13
VII	317	18	283	38	339	24
VIII	549	57	488	72	625	91
IX	324	31	312	22	343	35
XIV	87	17	84	8	94	19
X	215	50	207	44	235	67
XI	34	10	39	5	57	9
XII	74	3	68	1	75	5
XIII	2118	207	1957	153	2146	187

Cuadro 48: N° de noticias y n° de casos mensuales sobre tránsito
Fuente: Elaboración propia

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	4	11	14	14	6	14
I	6	10	10	26	1	19
II	30	14	56	38	22	11
III	15	7	26	16	22	2
IV	10	3	10	25	5	9
V	61	34	238	82	131	59
VI	23	6	18	19	19	8
VII	12	10	15	48	12	20
VIII	25	80	115	100	25	52
IX	6	12	25	37	11	21
XIV	9	19	55	52	15	33
X	9	25	55	46	15	27
XI	2	4	3	13	1	1
XII	4	2	34	8	2	1
XIII	289	161	684	346	314	183

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	17	13	5	6	5	4
I	3	9	5	16	3	7
II	63	8	56	10	36	10
III	26	5	16	7	24	3
IV	10	3	5	4	5	2
V	142	78	83	50	75	36
VI	48	8	16	5	10	2
VII	13	19	8	20	11	6
VIII	61	58	33	45	23	36
IX	21	26	12	29	7	14
XIV	30	19	19	15	13	15
X	30	29	19	21	13	20
XI	2	8	2	0	0	2
XII	13	3	2	1	4	2
XIII	359	257	137	196	214	132

Cuadro 49: N° de noticias y n° de casos mensuales sobre disturbios
Fuente: Elaboración propia

Región	Julio		Agosto		Septiembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	2	5	8	6	6	4
I	14	3	9	1	10	1
II	14	8	14	14	11	5
III	14	8	8	2	4	0
IV	19	2	13	0	19	1
V	54	31	60	12	60	33
VI	28	9	44	10	30	1
VII	41	9	33	11	36	10
VIII	86	38	84	62	68	25
IX	42	14	79	17	51	15
XIV	60	14	64	12	48	18
X	28	32	23	34	29	32
XI	11	1	18	0	8	1
XII	10	2	11	0	14	4
XIII	136	58	136	69	148	50

Región	Octubre		Noviembre		Diciembre	
	N° Casos	N° Noticias	N° Casos	N° Noticias	N° Casos	N° Noticias
XV	7	7	6	4	12	11
I	16	5	17	6	9	3
II	18	3	20	4	22	7
III	15	4	6	3	8	5
IV	24	2	23	0	17	3
V	73	28	70	20	85	28
VI	39	4	51	4	50	13
VII	44	8	46	21	67	29
VIII	75	31	92	45	172	59
IX	41	9	77	25	120	22
XIV	63	7	42	7	57	28
X	19	30	20	23	26	36
XI	15	3	10	0	13	3
XII	18	1	20	0	18	48
XIII	156	88	169	90	143	122

Cuadro 50: N° de noticias y n° de casos mensuales sobre incendios
Fuente: Elaboración propia

7.2. Estudio de modelos de regresión lineal por temática

7.2.1. Modelos de regresión lineal para la temática drogas

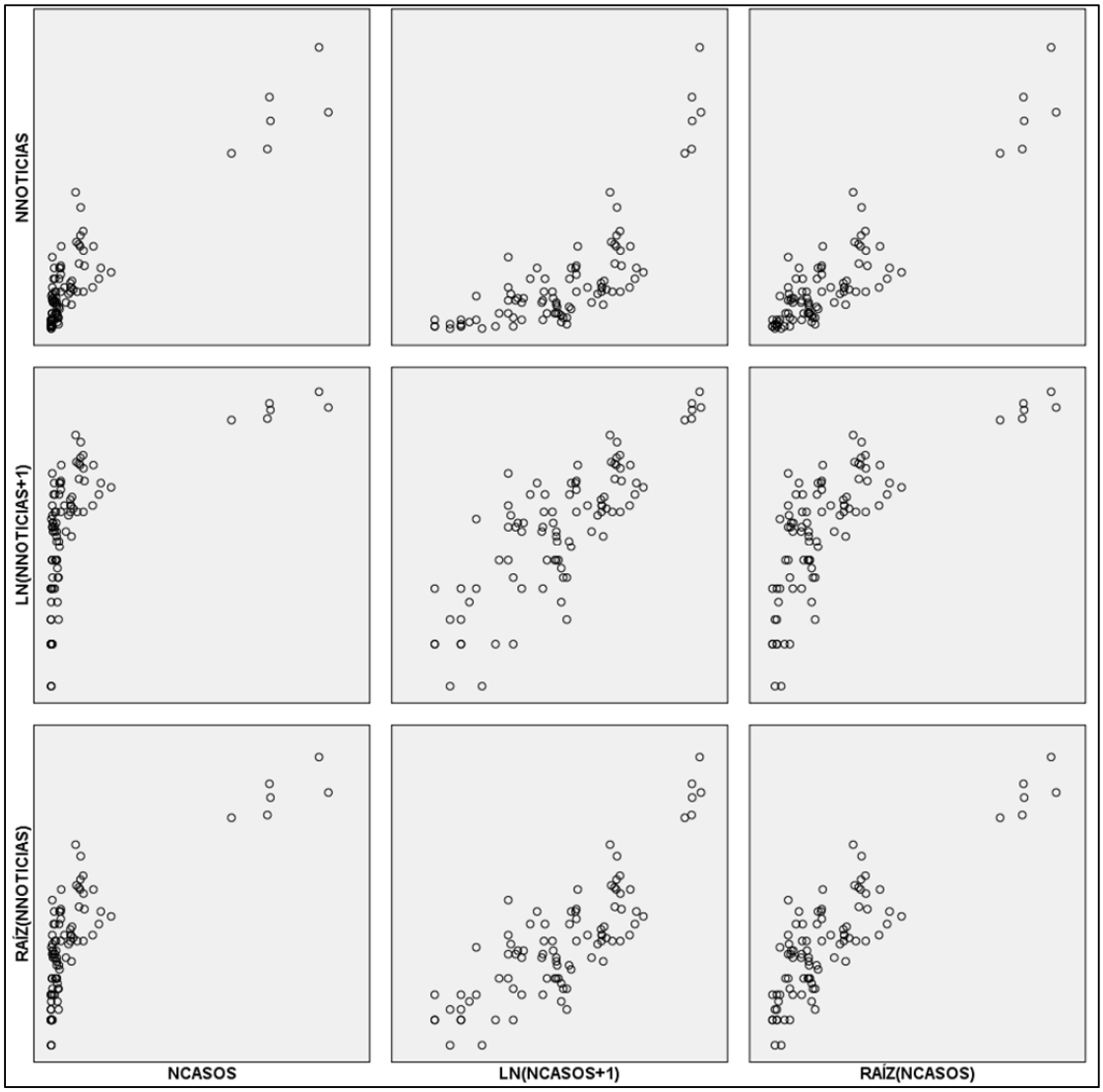

Figura 75: Diagrama de dispersión múltiple para la temática drogas
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 75, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
Raíz(N° Casos)	N° Noticias	no cumple	no cumple	cumple
Ln(N° Casos+1)	Ln(N° Noticias+1)	no cumple	no cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	no cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	cumple	cumple	cumple

Cuadro 51: Evaluación supuestos básicos regresión lineal, temática drogas
Fuente: Elaboración propia

Dentro de los modelos evaluados el único modelo de regresión lineal que cumple con los supuestos básicos de los errores es de la forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	340,706	1	340,706	230,244	0
Residual	130,219	88	1,48		
Total	470,925	89			

Cuadro 52: Tabla Anova del modelo de regresión seleccionado, temática drogas
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados			t	Sig.
		Error típ.			
Constante	2,069	0,186		11,119	0
Raíz(N°Casos)	0,19	0,013	0,851	15,174	0

Cuadro 53: Coeficientes del modelo de regresión seleccionado, temática drogas
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	2,063	
Normalidad	Shapiro-Wilk	0,980	0,186
Homocedasticidad	White	1,878	0,390

Cuadro 54: Supuestos del modelo seleccionado, temática drogas
Fuente: Elaboración propia

7.2.2. Modelos de regresión lineal para la temática robos

Figura 76: Diagrama de dispersión múltiple para la temática robos
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 76, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
Raíz(N° Casos)	N° Noticias	no cumple	no cumple	cumple
Ln(N° Casos+1)	Ln(N° Noticias+1)	no cumple	cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	no cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	cumple	cumple	cumple

Cuadro 55: Evaluación supuestos básicos regresión lineal, temática robos
Fuente: Elaboración propia

Dentro de los modelos evaluados el único modelo de regresión lineal que cumple con los supuestos básicos de los errores es de la forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	864,726	1	864,726	486,454	0
Residual	156,43	88	1,778		
Total	1021,155	89			

Cuadro 56: Tabla Anova del modelo de regresión seleccionado,, temática robos
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados		Coeficientes Estandarizados	t	Sig.
	β	Error típ.	β		
(Constante)	1,217	0,240		5,077	0,000
Raíz(N°Casos)	0,122	0,006	0,920	22,056	0,000

Cuadro 57: Coeficientes del modelo de regresión seleccionado, temática robos
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	1,883	
Normalidad	Shapiro-Wilk	0,985	0,419
Homocedasticidad	White	1,181	0,554

Cuadro 58: Supuestos del modelo seleccionado, temática robos
Fuente: Elaboración propia

7.2.3. Modelos de regresión lineal para la temática delitos sexuales

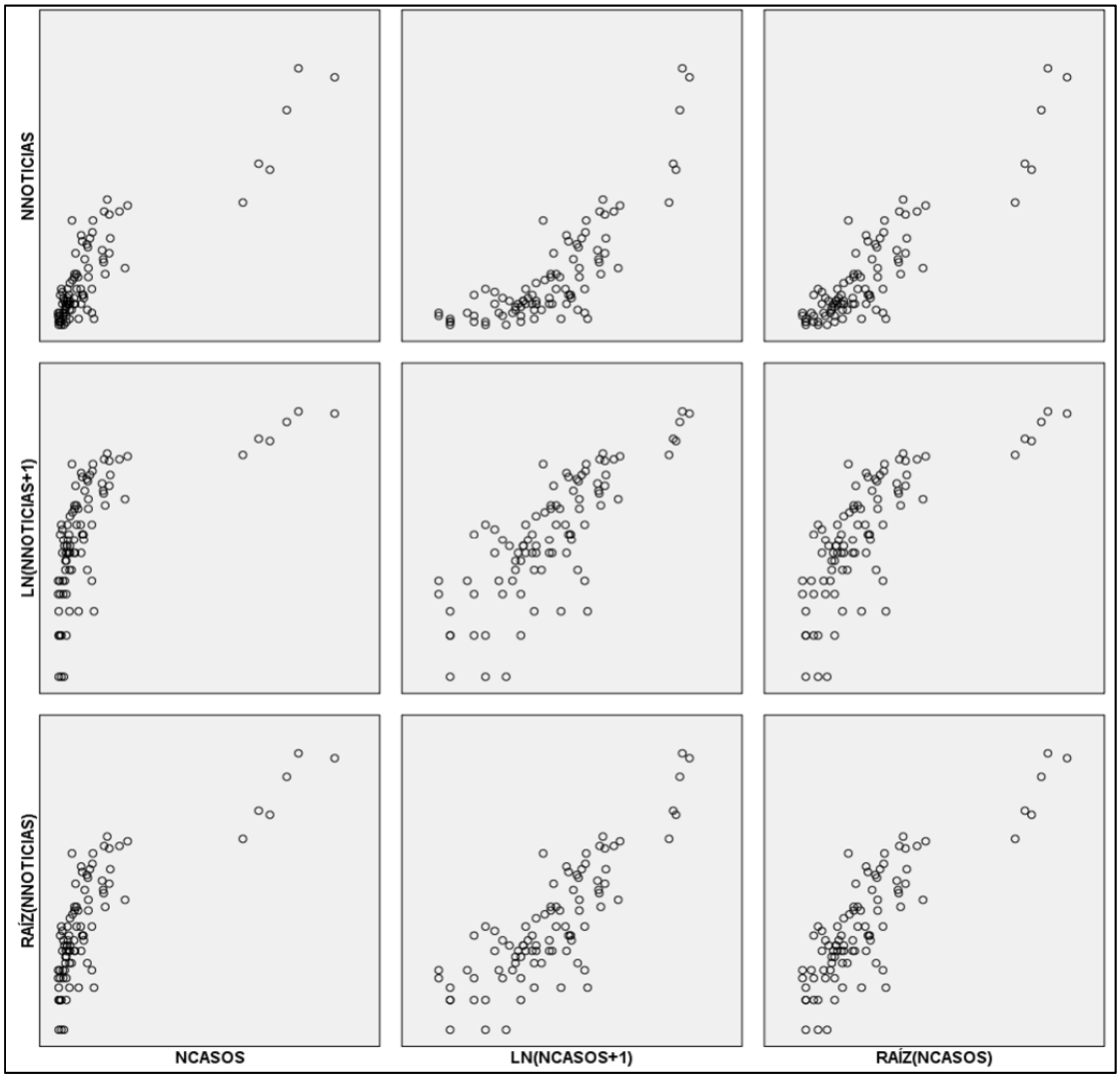

Figura 77: Diagrama de dispersión múltiple para la temática delitos sexuales
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 77, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
Raíz(N° Casos)	N° Noticias	cumple	no cumple	cumple
Ln(N° Casos+1)	Ln(N° Noticias+1)	no cumple	cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	cumple	cumple	cumple

Cuadro 59: Evaluación supuestos básicos regresión lineal, temática delitos sexuales
Fuente: Elaboración propia

En este caso se obtiene que dos modelos cumplen con todos los supuestos, por lo cual se selecciona al modelo que presenta un mayor R^2-ajustado:

Variable independiente	Variable dependiente	R^2-ajustado
Ln(N° Casos+1)	Raíz(N° Noticias)	0,688769
Raíz(N° Casos)	Raíz(N° Noticias)	0,699687

Cuadro 60: Comparación R^2-ajustado de distintos modelos, temática delitos sexuales
Fuente: Elaboración propia

Dentro de los modelos evaluados se selecciona el modelo de la forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	238,237	1	238,237	205,027	0
Residual	102,254	88	1,162		
Total	340,491	89			

Cuadro 61: Tabla Anova del modelo de regresión seleccionado, temática delitos sexuales
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados		Coeficientes Estandarizados	t	Sig.
	β	Error típ.	β		
(Constante)	1,054	0,208		5,059	0,000
Raíz(N°Casos)	0,362	0,025	0,836	14,319	0,000

Cuadro 62: Coeficientes del modelo de regresión seleccionado, temática delitos sexuales
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	2,246	
Normalidad	Shapiro-Wilk	0,993	0,907
Homocedasticidad	White	2,464	0,291

Cuadro 63: Supuestos del modelo seleccionado, temática delitos sexuales
Fuente: Elaboración propia

7.2.4. Modelos de regresión lineal para la temática homicidios

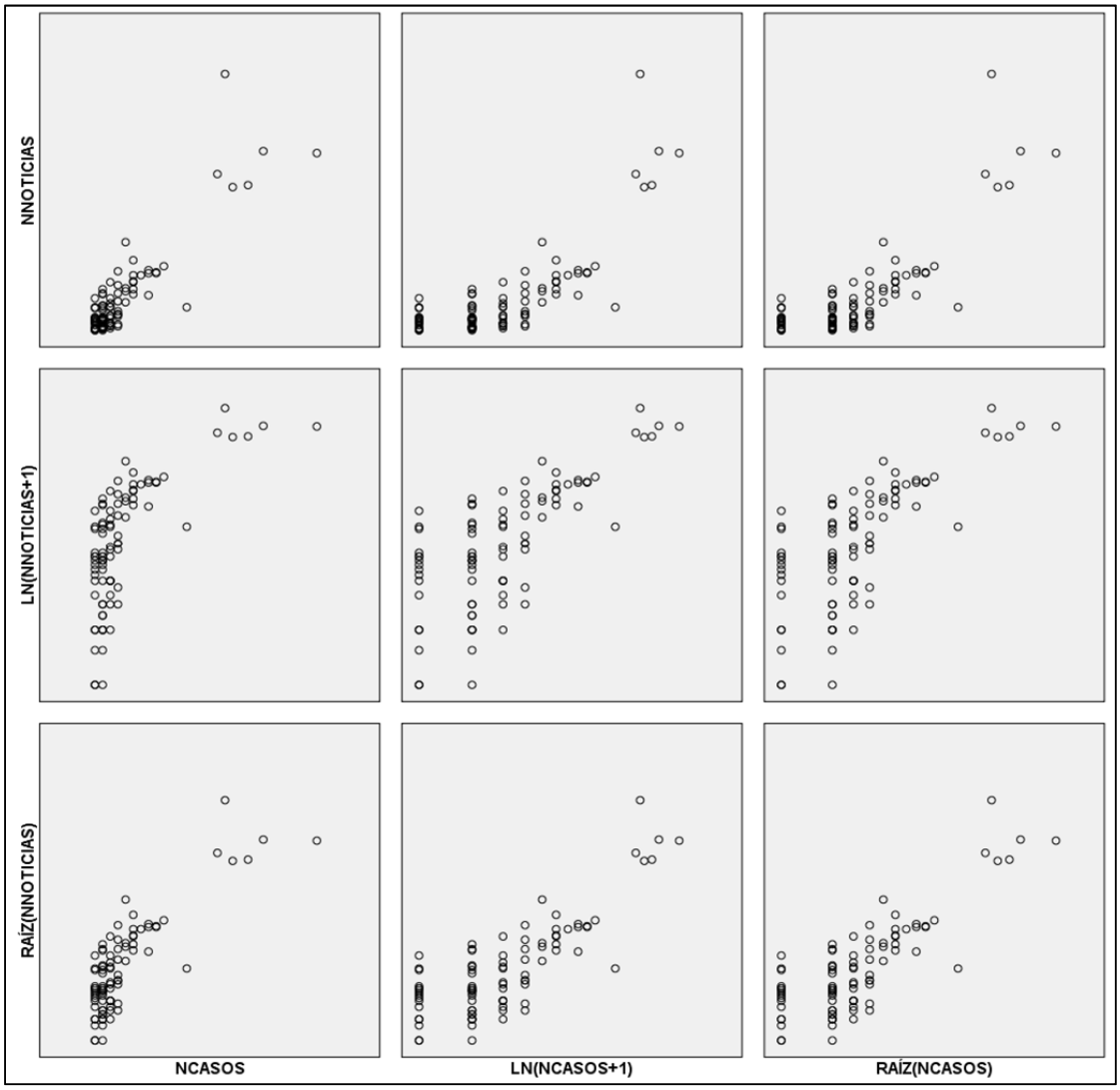

Figura 78: Diagrama de dispersión múltiple para la temática homicidios
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 78, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
Ln(N° Casos+1)	N° Noticias	no cumple	no cumple	cumple
Raíz(N° Casos)	N° Noticias	no cumple	no cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	no cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	cumple	cumple	cumple

Cuadro 64: Evaluación supuestos básicos regresión lineal, temática homicidios
Fuente: Elaboración propia

Dentro de los modelos evaluados el único modelo de regresión lineal que cumple con los supuestos básicos de los errores es de la forma:

$$\sqrt{(N° \ Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \ Casos)} + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	570,099	1	570,099	167,879	0
Residual	298,839	88	3,396		
Total	868,938	89			

Cuadro 65: Tabla Anova del modelo de regresión seleccionado,, temática homicidios
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados		Coeficientes Estandarizados	t	Sig.
	β	Error típ.	β		
(Constante)	1,539	0,317		4,862	0,000
Raíz(N°Casos)	2,193	0,169	0,810	12,957	0,000

Cuadro 66: Coeficientes del modelo de regresión seleccionado, temática homicidios
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	1,773	
Normalidad	Shapiro-Wilk	0,995	0,986
Homocedasticidad	White	1,426	0,490

Cuadro 67: Supuestos del modelo seleccionado, temática homicidios
Fuente: Elaboración propia

7.2.5. Modelos de regresión lineal para la temática tránsito

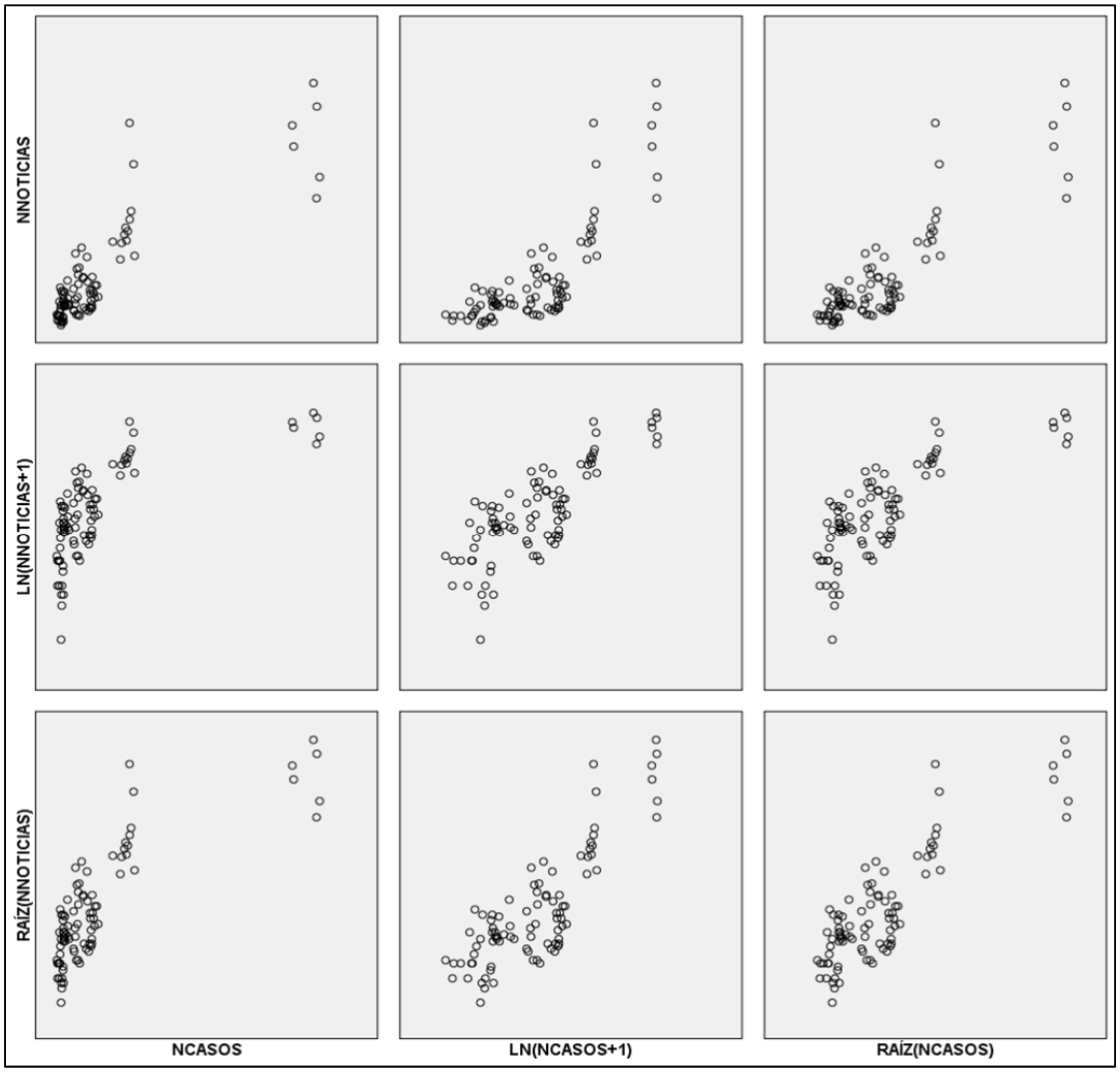

Figura 79: Diagrama de dispersión múltiple para la temática tránsito
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 79, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
Ln(N° Casos+1)	Ln(N° Noticias +1)	no cumple	cumple	cumple
Raíz(N° Casos)	Ln(N° Noticias +1)	cumple	cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	cumple	cumple	cumple

Cuadro 68: Evaluación supuestos básicos regresión lineal, temática tránsito
Fuente: Elaboración propia

En este caso se obtiene que dos modelos cumplen con todos los supuestos, por lo cual se selecciona el modelo que presenta un mayor R^2-ajustado:

Variable independiente	Variable dependiente	R^2-ajustado
Raíz(N° Casos)	Ln(N° Noticias +1)	0,5898
Ln(N° Casos+1)	Raíz(N° Noticias)	0,6839
Raíz(N° Casos)	Raíz(N° Noticias)	0,7276

Cuadro 69: Comparación R^2-ajustado de distintos modelos, temática tránsito

Dentro de los modelos evaluados se selecciona el modelo de la forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * \sqrt{(N° \, Casos)} + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	543,826	1	543,826	238,819	0,000
Residual	200,389	88	2,277		
Total	744,215	89			

Cuadro 70: Tabla Anova del modelo de regresión seleccionado, temática tránsito
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados		Coeficientes Estandarizados	t	Sig.
	β	Error típ.	β		
(Constante)	1,513	0,314		4,822	0,000
Raíz(N°Casos)	0,258	0,017	0,855	15,454	0,000

Cuadro 71: Coeficientes del modelo de regresión seleccionado, temática tránsito
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	2,187	
Normalidad	Shapiro-Wilk	0,978	0,123
Homocedasticidad	White	5,522	0,063

Cuadro 72: Supuestos del modelo seleccionado, temática tránsito
Fuente: Elaboración propia

7.2.6. Modelos de regresión lineal para la temática disturbios

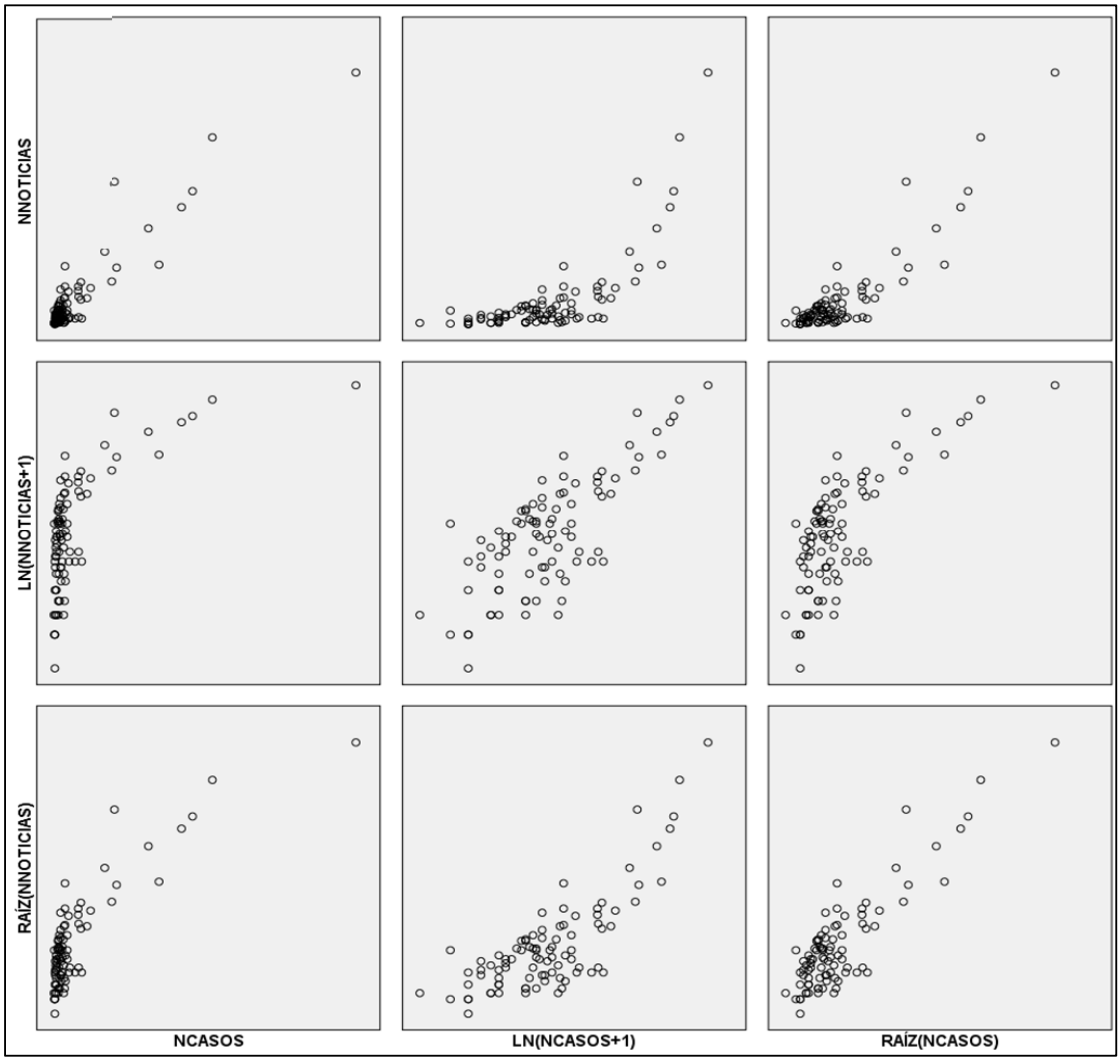

Figura 80: Diagrama de dispersión múltiple para la temática disturbios
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 80, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
N° Casos	N° Noticias	no cumple	no cumple	cumple
N° Casos	Ln(N° Noticias)	no cumple	no cumple	cumple
Ln(N° Casos+1)	Ln(N° Noticias+1)	cumple	cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	no cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	cumple	no cumple	cumple

Cuadro 73: Evaluación supuestos básicos regresión lineal, temática disturbios
Fuente: Elaboración propia

Dentro de los modelos evaluados el único modelo de regresión lineal que cumple con los supuestos básicos de los errores es de la forma:

$$Ln(N° \, Noticias + 1) = \beta_0 + \beta_1 * Ln(N° \, Casos + 1) + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	73,943	1	73,943	124,418	0
Residual	52,299	88	0,594		
Total	126,242	89			

Cuadro 74: Tabla Anova del modelo de regresión seleccionado, temática disturbios
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados		Coeficientes Estandarizados	t	Sig.
	β	Error típ.	β		
(Constante)	0,710	0,203		3,489	0,001
Ln(N°Casos+1)	0,713	0,064	0,765	11,154	0,000

Cuadro 75: Coeficientes del modelo de regresión seleccionado, temática disturbios
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	1,960	
Normalidad	Jarque-Vera	3,509	0,172
Homocedasticidad	White	1,939	0,379

Cuadro 76: Supuestos del modelo seleccionado, temática disturbios
Fuente: Elaboración propia

7.2.7. Modelos de regresión lineal para la temática incendios

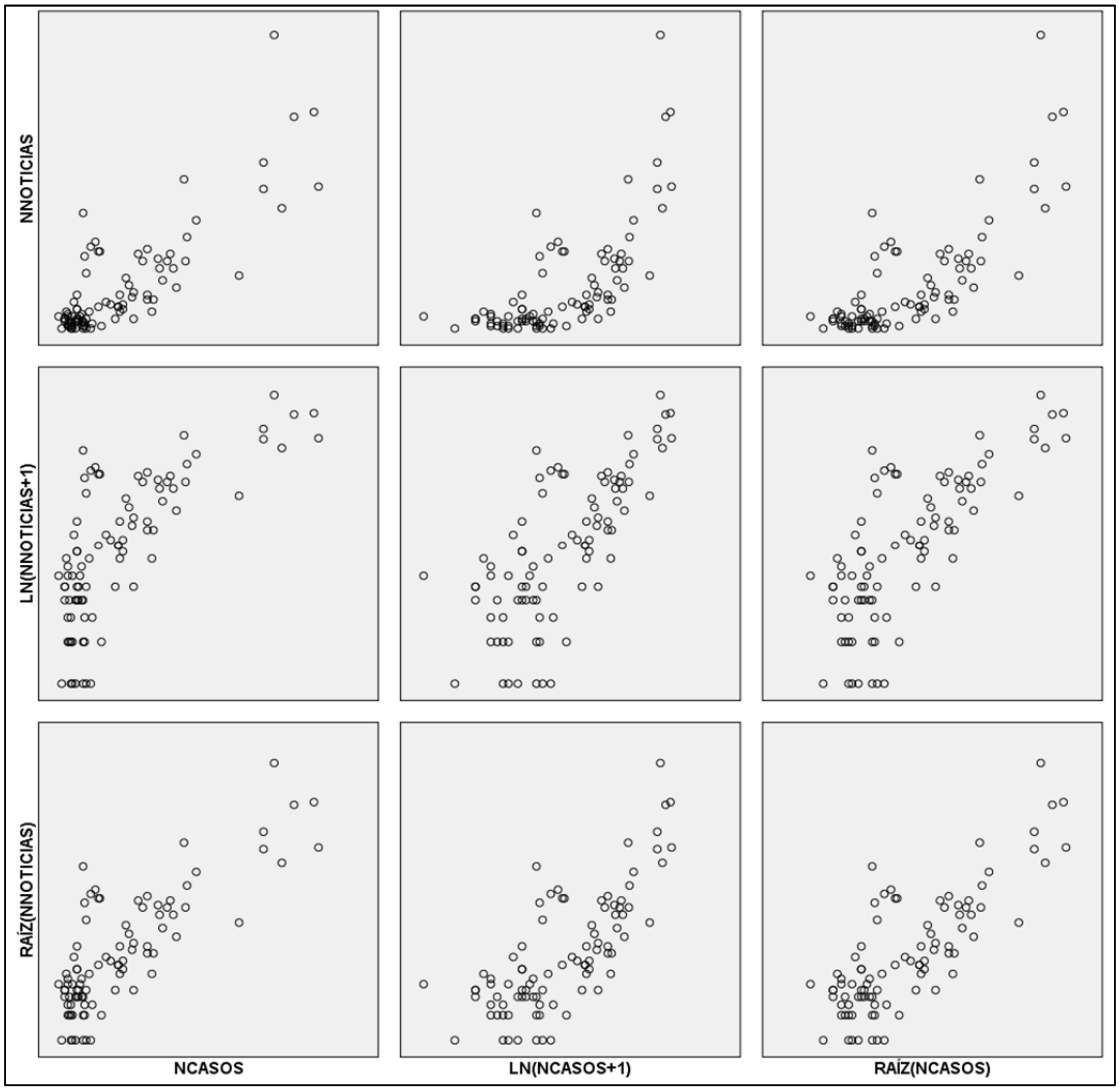

Figura 81: Diagrama de dispersión múltiple para la temática incendios
Fuente: Elaboración propia

A partir de los diagramas de dispersión presentados en la Figura 81, se seleccionan aquellos pares de variables que sugieren una relación lineal aproximada entre ellas y se verifican los supuestos básicos que deben cumplir los residuales de la regresión lineal:

Variable independiente	Variable dependiente	Test de normalidad	Test de homo-cedasticidad	Test de independencia
N° Casos	N° Noticias	no cumple	no cumple	cumple
N° Casos	Raíz(N° Noticias)	no cumple	cumple	cumple
Ln(N° Casos+1)	Raíz(N° Noticias)	cumple	cumple	cumple
Raíz(N° Casos)	Raíz(N° Noticias)	no cumple	cumple	cumple

Cuadro 77: Evaluación supuestos básicos regresión lineal, temática incendios
Fuente: Elaboración propia

Dentro de los modelos evaluados el único modelo de regresión lineal que cumple con los supuestos básicos de los errores es de la forma:

$$\sqrt{(N° \, Noticias)} = \beta_0 + \beta_1 * Ln(N° \, Casos + 1) + \varepsilon$$

Modelo	Suma de Cuadrados	gl	Media Cuadrática	F	Sig.
Regresión	285,013	1	285,013	113,062	0,000
Residual	221,836	88	2,521		
Total	506,849	89			

Cuadro 78: Tabla Anova del modelo de regresión seleccionado, temática incendios
Fuente: Elaboración propia

Modelo	Coeficientes no Estandarizados		Coeficientes Estandarizados	t	Sig.
	β	Error típ.	β		
(Constante)	-3,133	0,642		-4,880	0,000
Ln(N°Casos+1)	1,955	0,184	0,750	10,633	0,000

Cuadro 79: Coeficientes del modelo de regresión seleccionado, temática incendios
Fuente: Elaboración propia

Supuesto a Verificar	Test	Estadístico	p-valor
Independencia	Durbin-Watson	1,834	
Normalidad	Shapiro-Wilk	0,977	0,109
Homocedasticidad	White	2,043	0,360

Cuadro 80: Supuestos del modelo seleccionado, temática incendios
Fuente: Elaboración propia

8. Apéndice

8.1. Regresión lineal simple

El análisis de regresión [34] estudia la dependencia de una variable (variable dependiente) respecto de una o más variables (variables independientes o explicativas) cuyo objetivo es estimar o predecir la media de la variable dependiente en función de los valores conocidos de la variable independiente, lo que no implica necesariamente causalidad entre las variables. La regresión lineal simple es el análisis de regresión más sencillo posible, es decir, la regresión con dos variables, en la cual la variable dependiente se relaciona con una sola variable explicativa.

La relación funcional entre la verdadera media de Y_i, $E(Y_i)$, y X_i es la ecuación de una recta: $E(Y_i) = \beta_0 + \beta_1 X_i$, donde la β_0 es el intercepto (punto donde la recta intercepta el eje Y) y β_1 es la pendiente de la recta. Considerando la desviación de una observación Y_i alrededor de su valor esperado, el modelo queda de la siguiente forma: $Y_i = \beta_0 + \beta_1 X_i + u_i$. Las observaciones en la variable dependiente son asumidas como aleatorias, las observaciones de la variable independiente son asumidas sin error (constantes conocidas) y u_i es una variable aleatoria no observable.

El objetivo principal del análisis de regresión es estimar los parámetros del modelo $Y_i = \beta_0 + \beta_1 X_i + u_i$.

El método de los mínimos cuadrados ordinarios (MCO) selecciona como estimación de la recta de regresión poblacional aquella para la cual la suma cuadrada de los residuos es menor. Para el caso de regresión lineal simple, se tiene que buscar $\widehat{\beta_0}$ y $\widehat{\beta_1}$, las estimaciones numéricas de los parámetros β_0 y β_1 respectivamente, tales que minimicen la suma cuadrada de los residuos para una muestra dada:

$$\boldsymbol{min} \sum_{\boldsymbol{i}}^{\boldsymbol{n}} \boldsymbol{u_i^2} = \boldsymbol{min} \sum_{i}^{n} \left(Yi - \hat{Y}\right)^2 = \boldsymbol{min} \sum_{i}^{n} \left(Yi - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

donde $\boldsymbol{u_i} = Yi - \hat{Y}$ es el residuo observado de la i-ésima observación y n es el número de elementos de la muestra. La minimización se realiza usando las condiciones de primer orden, a través de lo cual se obtiene:

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X} \qquad\qquad \widehat{\beta_1} = \frac{\sum(Xi-\bar{X})(Yi-\bar{Y})}{\sum(Xi-\bar{X})^2}$$

Los estimadores obtenidos mediante este método se conocen como estimadores de mínimos cuadrados [68], que se caracterizan por ser insesgados, tener varianza mínima y presentar consistencia (los estimadores convergen a sus verdaderos valores a medida que aumenta el tamaño de la muestra).

Los supuestos básicos del modelo de regresión lineal son [34]:

- Linealidad: el modelo de regresión es lineal en los parámetros, aunque puede o no ser lineal en las variables. En general este supuesto puede ser revisado gráficamente a través del diagrama de dispersión de las variables.
- Normalidad: El modelo clásico de regresión lineal normal supone u_i está normalmente distribuida con media cero y varianza constante. Entre las pruebas para verificar este supuesto se encuentran el test de Shapiro-Wilk y el test de Jarque-Vera.
- Homocedasticidad o varianza constante de ui: la varianza del término de error es la misma sin importar el valor de la variable independiente. Para analizar este supuesto se realiza la prueba general de heteroscedasticidad de White.
- Independencia: no hay autocorrelación entre los errores, es decir, el error asociado a una observación es independiente del correspondiente a cualquier otra observación. Para estudiar este supuesto se realizan pruebas como el test d de Durbin-Watson.

Para determinar si el modelo estimado explica satisfactoriamente la muestra en estudio primero se debe analizar la variación total de la variable dependiente para toda la muestra y se considera la suma de todas las variaciones al cuadrado:

$$\sum(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}} = \sum(\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}} + \sum(\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}}_{\boldsymbol{i}})^{\boldsymbol{2}}$$

$$\sum(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}} = \sum(\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}} + \sum \boldsymbol{u_i^2}$$

$$\boldsymbol{SCT} = \boldsymbol{SCE} + \boldsymbol{SCR}$$

SCT es la suma de cuadrados totales y representa la variabilidad total en la muestra de la variable dependiente alrededor de su media. SCE es la suma de cuadrados de la regresión y representa la variabilidad explicada por la regresión. SCR es la suma de cuadrados de los errores y representa la variabilidad que permanece sin explicar debido al error.

El coeficiente de determinación, r^2, mide la proporción de la variabilidad de la variable dependiente explicada por su relación lineal con la variable independiente, indicando cuan bien se ajusta la línea de regresión muestral a los datos [34]. El valor de r^2, es una medida sin unidades, correspondiente al cuadrado del coeficiente de correlación lineal (que mide el grado de asociación entre dos variables), está comprendido en el intervalo [0,1]. A mayor valor de r^2, mayor será el grado de ajuste y r^2=0 implica que no existe relación lineal. Este coeficiente queda definido por:

$$r^2 = \frac{\sum(\widehat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})} = \frac{SCE}{SCT} = 1 - \frac{\sum(\widehat{u}_i)^2}{\sum(Y_i - \overline{Y})} = 1 - \frac{SCR}{SCT}$$

Una forma de resumir la descomposición de la variabilidad de la variable dependiente es emplear la tabla ANOVA, la cual concentra importante información del modelo y a través de la cual se puede determinar si el modelo de regresión lineal sirve para relacionar las variables (realizando un estudio del estadístico F).

Modelo	**Suma de Cuadrados**	**gl**	**Media Cuadrática**	**Test F**
Regresión	SCE	1	MCE=SCE/1	F=MCE/MCR
Residual	SCR	N-2	MCR=SCR/(N-2)	
Total	SCT	N-1	SCT(N-1)	

En análisis de regresiones lineales muchas veces se utiliza la transformación de las variables en estudio, tanto de la variable dependiente como de las independientes, debido a que existen diversas razones que la justifican, entre las que destacan [68]: resolver el problema de no normalidad en la distribución de los errores y/o presencia de heterogeneidad en la varianza de los errores y la simplificación de la relación existente entre la variable dependiente y la variable independiente. En el caso de que no se tiene una idea previa de la forma del modelo que relaciona la variable dependiente e independiente, se estudian empíricamente diferentes formas matemáticas que representen la relación de la forma más simple posible (idealmente lineal en los parámetros). Por otro lado, cuando se sabe previamente que la relación no es lineal en los parámetros, por ejemplo $Y = \alpha X^{\beta}$, se busca reexpresar el modelo en uno lineal en los parámetros, de la forma $Y^{*} = \alpha^{*} + \beta X^{*}$ (para el ejemplo anterior sería: $\ln(Y) = \ln(\alpha) + \beta \ln(X)$), que permita el uso de mínimos cuadrados ordinario para la estimación de los parámetros. Las relaciones más simples permiten entender de mejor forma las relaciones entre variables y las relaciones más simples corresponden a aquellas que tienen menos parámetros y que son lineales.