PROYECTOS FONDEF DE INVESTIGACIÓN Y DESARROLLO

PROGRAMA IDeA

INFORME DE AVANCE PARCIAL

CIENTÍFICO–TECNOLÓGICO

Nº 1

Título del Proyecto: Observatorio Digital de Delincuencia en Chile: Un sistema inteligente de apoyo a la industria automotriz chilena, en el robo de vehículos y accesorios.

Código del Proyecto: ID16I10222

Nombre Director(a) del Proyecto: Richard Weber Haas

Institución(es) Beneficiaria(s): Universidad de Chile, Pontificia Universidad Católica de Chile

Institución(es) Interesadas o Socias Contraparte: DERV-AACH (antes PROSE)

Nombre Ejecutivo(a) de FONDEF: Andrea Hinojosa Moreno

Fecha de emisión del informe: 29/09/2017

Período correspondiente a este informe: Entre el 03/01/2017 y el 25/09/2017

COMISIÓN NACIONAL DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA
MONEDA 1375 • FONO: 56 2 2365 44 00 • FAX: 56 2 2655 13 94 • SANTIAGO CHILE

INFORME DE AVANCE PARCIAL
PROGRAMA IDeA
ÍNDICE

1. PRESENTACIÓN DEL PROYECTO

En esta sección agregue el resumen, las hipótesis y los objetivos generales y específicos del proyecto. Puede copiar dicha información desde el proyecto original.

RESUMEN	El robo de vehículos es un tema importante en todo el mundo, teniendo costos tanto económicos como sociales. Este crimen trae una pérdida económica para el dueño del vehículo, pero a menudo trae incluso mayores costos para las víctimas y la sociedad, generando una percepción de inseguridad mayor. En países como Chile, donde el parque automotriz sigue aumentando, la aparición de este problema va a empeorar en el futuro si no se toman las medidas adecuadas. En países como Estados Unidos, entre otros, si bien se han tomado varias medidas para evitar el robo de vehículos y para encontrar vehículos robados, no existe un sistema inteligente que detecte los patrones existentes y emergentes en el robo de vehículos. Con este propósito, se propone un observatorio para apoyar a la industria automotriz chilena. Actualmente son once las compañías de seguros automotrices en Chile, las cuales fundaron PROSE Chile como su consultora en temas de seguridad. PROSE Chile está cooperando en este proyecto con su experiencia y con la base de datos de los siniestros reportados por los asegurados. El nivel tecnológico de las compañías de seguros en Chile hace que la alternativa de construir su propio sistema de monitoreo sea extremadamente cara. PROSE Chile tiene las plataformas tecnológicas para implementar un proyecto de este tipo, pero carece del capital humano altamente capacitado para su desarrollo, por lo que el proyecto propuesto hace viable esta iniciativa. La mayoría de las medidas para prevenir el robo de vehículos no utilizan tecnologías modernas, como la ciencia de datos y el análisis de redes sociales, parte fundamental del observatorio propuesto. Para ello, se propone consolidar la información disponible de diferentes fuentes, es decir, el registro de siniestros, mensajes en las redes sociales, así como noticias relacionadas con el robo de vehículos, la cual se encuentra en distintos tipos de datos, tales como numéricos, de texto y no estructurados, por lo que exigen de investigación aplicada para lograr el sistema propuesto. Con este observatorio implementado se espera una reducción del robo de vehículos en Chile del 5% debido a la aplicación de medidas preventivas adecuadas por las compañías de seguros, los dueños de vehículos, y las policías. Esta reducción resulta en un ahorro de CLP $ 3.000 millones / año (aprox. USD $4.500.450 / año) sólo para vehículos asegurados, los que representan un tercio del total. Múltiples actores se verán beneficiados por este observatorio. En primer lugar las compañías de seguros que recibirán menos siniestros por la reducción del robo de vehículos. Esto se traduciría en una

	reducción en las primas de los seguros que favorece a los dueños de vehículos. Los organismos de seguridad van a mejorar su información para basar sus estrategias en la prevención del delito. La sociedad también ganará debido a una mejor prevención del delito que lleva a una mayor percepción de seguridad. Disclaimer (Cambio de nombre de entidad interesada): El Directorio de Prose Chile, a comienzos de 2017, explicó a la AACH que ellos daban por concluidos en forma exitosa los objetivos para los cuales se habían constituido y que tenían relación con promover medidas para combatir los robos de vehículos, una institucionalidad y políticas para ello, razón por la cual entrarían en una etapa de término de sus actividades. Esto implicaba descontinuar la recopilación de estadísticas que utilizaban como soporte para sus objetivos. La AACH, aunque no hará las funciones que hacía Prose, estimó que las estadísticas de robos que llevaba Prose-Chile, debieran tener una continuidad por el valor que tienen, y por ello decidió formar un departamento que se encargase de esa parte. Esta nueva área se denomina “Departamento de Estadísticas de Robos de Vehículos (DERV)”, y operará de manera similar al DIS, con foco exclusivo en el robo de vehículos, y bajo la supervisión de la actual Gerencia de Proyectos de la AACH.
HIPÓTESIS	**H1:** Los robos de vehículos y accesorios **no son eventos aleatorios**, y siguen **patrones detectables** a través de técnicas avanzadas de *data mining*. **H2:** Los **métodos de aprendizaje y visualización** permiten revelar de forma efectiva y eficaz los **patrones**, mejorando la experiencia del usuario.
OBJETIVO GENERAL	Desarrollar un Observatorio Digital del robo de vehículos y accesorios en Chile, capaz de caracterizar los modi operandi de los delincuentes, y su evolución, provenientes de las denuncias y fuentes de información en línea, a través de técnicas avanzadas de Data Mining.
OBJETIVO ESPECÍFICO 1	Caracterizar los modi operandi de los delincuentes

OBJETIVO ESPECÍFICO 2	Detectar patrones en los robos.
OBJETIVO ESPECÍFICO 3	Obtener una herramienta de visualización de patrones de robo y presentación de indicadores relevantes para el usuario

2. EXIGENCIAS Y RECOMENDACIONES PREVIAS

Indique las exigencias y recomendaciones del Comité de Área, provenientes de la última presentación de avance. Además señale cuándo y cómo fueron resueltas.
Si este es el primer informe de avance, refiérase a las exigencias y recomendaciones del acta de adjudicación.

Exigencia 1:
Indicar cómo atacarán que la componente de la información a utilizar, no sea sesgada a lo provisto por las aseguradoras, lo que podrá dar un resultado sesgado de acción y siniestros.
Esto ya que existe un sesgo, que es la información que no manejan las aseguradoras. Se espera que el proyecto justifique este sesgo y plantee una solución al momento de ejecutar el proyecto.

Respuesta:
Para mitigar el efecto de este sesgo agregamos redes sociales y noticias como fuentes para alimentar el observatorio. Además tenemos una carta de apoyo de carabineros y solicitaremos datos de denuncias que manejan en Carabineros.

Exigencia 2:
Incorporar expertos del área de dominio del problema (robo de vehículos y sus accesorios, comportamiento social, etc.) Un experto social en robos o delincuencia.

Respuesta:
Con la participación de un profesional de la empresa Prose Chile ya tenemos un experto en el área. Además invitaremos a expertos del CESC (Centro de Estudios de Seguridad Ciudadana http://www.cesc.uchile.cl/) con el cual hemos tenido interacciones durante los años pasados.

Exigencia 3:
Se solicita enviar las autorizaciones de los organismos correspondientes que garanticen la protección de la información que se utilizará, y un documento donde se indiquen las mediadas de resguardo que se adoptarán.

Respuesta:
Está listo.

Exigencia 4:
Revisión de planilla de costos e ingreso de presupuesto inicial en plataforma, en función de requerimientos de las bases y comentarios del panel evaluador.

Respuesta:
Se hizo.

Exigencia 5:
Revisión de resultados y de su caracterización en la plataforma de S+C, incorporación de hitos correspondientes, programación de fechas de logro tanto de los hitos como de los resultados.

Respuesta:
Se hizo.

3. DESCRIPCIÓN SINTÉTICA DEL AVANCE DEL PROYECTO

Resultado Producción / Hitos	Fecha de logro comprometida en plataforma	Fecha de logro estimada	Porcentaje de avance a la fecha
1. Resultado Producción N°1 Caracterización preliminar y análisis de patrones de los datos de robos de vehículos y accesorios desde las fuentes de AACH-DERV (antes: PROSE Chile), Twitter y medios noticiosos en línea."	30-09-2017	30-09-2017	95%
1.1 Hito 1 "Caracterización del robo desde las fuentes de AACH-DERV."	30-09-2017	30-09-2017	100%
1.2 Hito 2 "Caracterización del robo desde Twitter"	30-09-2017	30-09-2017	100%
1.3 Hito 3 "Caracterización del robo desde medios noticiosos en línea"	30-09-2017	30-09-2017	80%

4. MARCO CONCEPTUAL Y ESTADO DEL ARTE

4.1. Estudios sobre robo de autos

En la literatura de criminología, en general, los robos vehiculares no han recibido mucho interés investigativo en comparación a otros crímenes. Si bien este tipo de delito no es diferente a otros en el robo de propiedad, si es diferente en su información.
Los hurtos son mayormente físicos. Donde el delincuente transgrede a la víctima de forma física y/o emocional, para su beneficio personal en la adquisición de bienes personales y tangibles, para el uso y usufructo de los mismos. Los robos de vehículos motorizados ocurren de forma imprevista, y son descubiertos difícilmente en el momento. Teniendo bajas tasas en la aprehensión delictual; y siendo mayoritariamente, descubiertos cuando el acto de beneficio personal del delincuente ya fue concretado.

4.1.1. Comportamiento del delincuente

Para entender el robo vehicular, es necesario tener en consideración las posibles causas –y explicación de las mismas –detrás del comportamiento del delincuente. Existen dos teorías que postulan una explicación a dichas causales dentro de este contexto. En primer lugar, está la Teoría de Desorganización Social (Groves y Sampson, 1989) que hace referencia a la incapacidad de una comunidad de hacer efectivos los valores de sus miembros, y de mantener un control efectivo sobre sus conductas. En paralelo está The Routine Activity Theory (Cohen y Felson, 1979) la cual afirma que los crímenes ocurren

cuando existen objetivos atractivos, delincuentes motivados y ausencias en el control policial. Creando, por parte del delincuente, una decisión racional que maximiza sus ganancias, y minimiza las pérdidas asociadas (Barclay et al. 1996; Clarke, 1989).

4.1.2. Experiencias Internacionales

Considerando lo anterior, es claro que para distintos países existen escenarios distintos. Salario mínimo, tasas de desempleo, acceso a créditos para la compra/venta de vehículos motorizados; son solo algunas características que pueden incidir de manera distinta en los resultados, y actos delictuales; cometidos en cada uno de ellos.
Un estudio muestra que el robo de automóviles ha sido el delito más frecuente en Taiwán en el 2004 (Chen et al. 2006). Asimismo, en el caso de Israel, que desde 1994 ha tenido un crecimiento sistemático en las tasas delictuales hasta el año 1997 (Herzog, 2002). A diferencia de Gran Bretaña y Estados Unidos que tuvieron una tendencia decreciente en el mismo período (HEUNI, 1997; US Federal Bureau of Investigation, 1991).

4.1.3. Tipología del robo de vehículos motorizados

En criminología, dentro del contexto de robos de vehículos motorizados; se han desarrollado escasos intentos en desarrollar tipologías específicas sobre este tema (Clarke, 1991; Clarke y Harris, 1992; McCaghy et al. 1997). Para el presente informe, se tendrá en consideración los tipos más sintetizados, y simplificados; presentados por (Arizona Criminal Justice Commission, 2004) la cual expone tres categorías: (1) Robo con fines de lucro, (2) robo para asegurar transporte y (3) robo recreacional.

4.1.4. Análisis de trabajos seleccionados

The Key to Auto Theft: Emerging Methods of Auto Theft from the Offenders' Perspective: Los autores de este paper (Copes y Cherbonneau, 2006) realizan un estudio introspectivo sobre 2 grupos de control compuesto por reclusos con distinto nivel de peligrosidad, pero que tienen en común el robo de vehículos motorizados. El primero de ellos con 42 individuos que cometieron al menos un robo de estas características, y el segundo; 12 personas encarceladas en prisiones de mediana seguridad en Louisiana. Se busca determinar los métodos (y metodología) preferencial que posee el delincuente a la hora de hurtar, caracterizando el trabajo en distintas componentes para poder concluir sobre qué factor es más importante. Entre los factores evaluados se encuentran: (1) características: demográficas, psicológicas, físicas, adictivas; entre otras. (2) Alert opportunism, (3) búsqueda activa de oportunidades, entre otros factores. Los resultados sugieren las componentes no son solo oportunistas. Si no, presentan cierto grado de razonamiento al momento de delinquir (Groves y Sampson 1989; Cohen y Felson 1979) y que se ha visto en la obligación de evolucionar, producto de los cambios estructurales que tienen periódicamente los diseños vehiculares.

Empirical Analysis of Motor Vehicle Theft in Israel, 1990-97: En este trabajo (Herzog, 2002) se busca presentar las principales características en el robo de vehículos motorizados. Contrastándolo con las categorías más amplias (crimen general o crimen de propiedad) y la relación de los datos, tanto con la población total; como con aquellos registrados en los robos vehiculares. Asimismo, se busca caracterizar el fenómeno en base

a las características propias de la base de datos a analizar. Se presenta un análisis estadístico sobre datos recopilados en el robo de vehículos motorizados dentro de Israel. Dicha información es tratada en dos niveles de análisis: (1) análisis según frecuencia relativa de la variable (característica) a analizar y (2) análisis estadístico mediante el uso de modelos de regresión logística. Filtrando de forma binaria a aquellos vehículos que fueron, o no, recuperados. Los resultados muestran que características tales como el año y tipo de vehículo; así como también la religión, edad y nacionalidad (israelí o palestina) del delincuente; son significativas al momento de caracterizar el fenómeno, atribuyendo el robo de vehículos motorizados a una combinación de influencias políticas y económicas, en la región.

Locations of Motor Vehicle Theft and Recovery: En esta investigación (Suresh y Tewksbury, 2013) se examinan localidades de riesgo (hot-spot) en el robo de vehículos motorizados dentro de la región de Louisville, Kentucky desde el 2004 al 2007. Con ello, se busca encontrar una relación entre aquellas zonas periféricas a la delincuencia y las tasas de robos correspondientes a cada una de ellas, junto con patrones que indiquen el desplazamiento mismo de los hot-spot. Se identifican características – y contexto – del lugar donde ocurren los siniestros a través de modelos de regresiones espaciales. Primero, observando la correlación espacio-tiempo de las zonas donde ocurren los robos de vehículos motorizados, y dónde éstos son recuperados. A través de mapas LISA (Local Indicators of Spatial Association) y estadísticos de Moran. Luego, se filtran y seleccionan las variables (características poblacionales y demográficas) que posean menor correlación para la caracterización de hot-spots. El estudio sugiero que, gracias a las regresiones espaciales; es posible confirmar la alta concentración de robos vehiculares; y la recuperación de los mismos, en barrios con indicadores de desorganización social (Groves y Sampson 1989).

Mining Criminal Database to Finding Investigation Clues – By Example of Stolen Automobiles Database: En este paper (Chen et al. 2006) se examina la información recopilada en Taiwan, sobre la delincuencia vehicular enfocada en el robo de vehículos motorizados. El objetivo de este estudio es presentar herramientas, e información vinculante; que no es visible de forma inmediata en la información. Encontrando patrones delictuales que puedan servir para el entendimiento de este tipo de crimen en particular. Para ello se utilizan técnicas y algoritmos propios del data mining. Posibilitando la captura de información relevante, y oculta; en los datos predispuestos a analizar. Para tales efectos se utilizan las técnicas de: (1) reglas de asociación, (2) clasificación, (3) regresiones predictivas y (4) data generalization & summarization-based characterization. Los resultados del estudio sugieren que en cuanto más pronto sea reportado el robo, más probable es su recuperación. En paralelo, Taiwán presenta altas demandas de vehículos marca Toyota. Por tanto, dicha marca es robada a menudo para desmontaje de sus piezas. Para la compra/venta de las mismas en el mercado negro.

Discovering Investigation Clues through Mining Criminal Databases: El autor de este paper (Chen, 2008) analiza 378.000 registros de robos vehiculares dentro de una ventana de tiempo de 11 años. Siguiendo la metodología expuesta en (Chen et al. 2006), propone un análisis enfocado en el tiempo transcurrido después de notificado el siniestro; y en cómo las características propias del vehículo pueden cooperar de forma simultánea en el

descubrimiento de información de utilidad. Dicho estudio muestra que los delincuentes prefieren delinquir en horarios nocturnos a altas horas de la madrugada. Entendiendo esto como el rango con mayor tasa delictiva aquel que fluctúa entre 4:00 AM y 8:00 AM. Adicionalmente a ello, se segmentan las locaciones de ocurrencia de los hechos; para ver a qué zona le corresponde la mayor ponderación. Observando que la mayor tasa delictiva está presente en las zonas: centro-norte de la región a analizar. Esto permite tener conocimiento de las ubicaciones, y cómo éstas pueden ser utilizadas – y clasificadas – como hot-spot. Para su respectivo análisis, y precaución en el control policial.

Implementation of Data Mining for Vehicle Theft Detection using Android Application: En este trabajo (Sharma, 2014) se busca crear una aplicación de celular, colaborativa; en donde los usuarios entrenen a la aplicación para que ésta pueda: mantener, rastrear y predecir a los criminales que hayan cometido robo con algún vehículo motorizado. Lo anterior es logrado gracias a la technique for sequential pattern data mining con el uso del algoritmo Apriori, y la **technique for frequent pattern data mining** con el uso del algoritmo FP Growth. Posteriormente se procesan los datos según la siguiente secuencia: (1) data collection, (2) pre-procesamiento de datos, (3) aplicación de algoritmos y (4) análisis de resultado. El estudio postula un mecanismo predictivo que permita descubrir a posibles criminales con la detección de patrones con el uso de eventos pasados.

4.2. Redes sociales

Alrededor de 30.000 autos son robados anualmente en Chile. El estudio de actividades humanas complejas como actos criminales, en particular robo de autos, requiere el análisis de múltiples fuentes de información. En este contexto, las redes sociales pueden aportar informacion valiosa en tiempo real que no puede encontrarse en otras fuentes. Una de las plataformas más utilizada y popular en Chile es Twitter, red social donde se comparten opiniones o declaraciones en tiempo real a través de mensajes con un máximo de 140 caracteres, llamados tweets. El objetivo de este trabajo es descubrir las relaciones existentes entre el robo de vehículos y las denuncias realizadas a través de redes sociales.

Twitter ha sido utilizado como fuente de información en múltiples trabajos con el objetivo de predecir eventos del mundo real. En el caso de eventos criminales el enfoque principal ha sido enriquecer modelos predictivos basados en registros históricos con información extraída de esta red social. Los patrones principales que han sido extraídos como indicadores de Twitter son Análisis de sentimientos y Modelado de tópicos. A continuación presentamos una selección de trabajos encontrados en la literatura.

4.2.1 Enfoque basado en análisis de sentimientos

El análisis de sentimientos es la tarea de extraer ”subjetividad vs objetividad” y clasificar ”positivo vs negativo vs neutro” la opinión expresada por el autor en un texto. Dado que los actos criminales provocan opiniones y reacciones en las personas, el estado de opinión relacionado a un lugar es una variable que es influencia y influye al mismo tiempo en otras variables entre ellas la probabilidad de que ocurra un acto criminal como el robo de un auto.

Crime pattern detection using online social media: En el trabajo de Bolla (2014) se

recolectaron un total de 100.000 tweets del 7 de julio al 27 de julio de 2014 de las diez ciudades más violentas y las diez menos violentas de los Estados Unidos. Los tweets fueron recolectados usando palabras claves asociados a actos criminales. Basándose en los trabajos presentados en Torii et al. (2011), Arsevska et al. (2016) y Chang et al. (2012), hicieron análisis de sentimientos sobre estos tweets. Finalmente realizaron análisis de correlación entre el nivel de criminalidad bas andose en los artículos de Forbes y patrones observados en Twitter. Primero realizan un análisis donde sólo tienen en cuenta el volumen de tweets, encontrando que las ciudades menos violentas de acuerdo al informe de Forbes presentan una menor proporción de tweets asociados a actos violentos con respecto al resto de las ciudades. Luego tienen en cuenta el análisis de sentimientos sobre los tweets, revelando que las ciudades con mayor índice de violencia presentan una mayor intensidad de sentimientos negativos, aunque algunas de las menos violentas también presentaron niveles altos de tweets con carga negativa.

Crime prediction using Twitter sentiment and weather: En este trabajo Sharef y Martin (2015) proponen un modelo de predicción de crımenes que indica cuándo y dónde ocurrirá un crimen. Este modelo está basado en KDE (kernel density estimation), utilizando variables climatológicas y análisis de sentimientos sobre tweets con regresión logística como algoritmo de predicción. KDE es una técnica que ajusta un espacio bi-dimensional de función de densidad de probabilidad. Recolectaron alrededor de 1.000.000 de tweets desde el 25 de diciembre del 2013 al 31 de enero del 2014 en el área de Chicago, Illinois. Como datos de evaluación se tuvo en cuenta el registro histórico de alrededor de 5.400 robos con su ubicación en el área de Chicago obtenidos a través del sitio web del departamento de policía. Para determinar la variable predicha por el modelo de predicción se dividió el área de Chicago en cuadrıculas de 200mx200m y se consideraron intervalos de tiempo de 6 horas para el cálculo de las variables correspondientes a cada cuadrícula. Los experimentos demostraron que la inclusión de información procedente de Twitter y de variables climatológicas mejora la capacidad predictiva del modelo.

4.2.2 Enfoque basado en tópicos

El modelado de tópicos se basa en modelos probabilísticos para descubrir la distribución de tópicos en una colección de documentos. Los trabajos presentados en esta sección utilizan el modelado de tópicos para extraer patrones que aporten información adicional a los sistemas de predicción de crimen.

Predicting crime using Twitter and kernel density estimation: Este trabajo (Gerber, 2014) tiene por objetivo crear un modelo para predecir la probabilidad de que cierto tipo de crimen ocurra en un lugar determinado para el día siguiente. Para eso los autores utilizan KDE (Kernel Density Estimation), al igual que en el trabajo de Sharef y Martin (2015), pero con datos diferentes. Para este estudio se recopilaron tweets geolocalizados por GPS desde el 1 enero al 21 de marzo del 2013, en la ciudad de Chicago. Además utilizaron datos históricos de crímenes en dicha ciudad en el mismo periodo de tiempo obtenidos a través del departamento de policıa. Consideraron 25 tipos de crímenes, entre los cuales se considera robo, daños criminales, violación del uso de armas, asaltos, robo de propiedad privada, robo de vehículos, homicidios, entre otros. Compararon dos tipos de modelos, el primero utiliza solamente registros históricos de crímenes, el segundo incluye los datos

provenientes de Twitter, que antes son procesados utilizando modelamiento de tópicos (Blei, Ng y Jordan, 2013; Blei, 2012). De los 25 tipos de crímenes considerados, 19 mostraron mejoras en la predicción cuando se añadieron los tópicos de Twitter al modelo KDE.

Automatic crime prediction using events extracted from Twitter posts: El objetivo principal de este trabajo (Wang, Gerber y Brown, 2012) es construir un modelo de predicción crímenes del tipo *hit-and-run* (atropellos con huída del conductor) basado en información procedente de la red social Twitter. Específicamente se busca predecir si un evento ocurrirá durante un día en particular o no. Para ello recolectaron tweets de agencias de noticias del área de Charlottesville, Virginia desde el 22 de Febrero al 21 de Octubre de 2011. Además consideraron en la construcción del modelo y como datos para la evaluación del modelo información procedente de registros históricos de crímenes de dicha ciudad. Para añadir la información procedente de Twitter al modelo de predicción se realizan dos pasos, primero Etiquetado de Rol Semántico (Gildea y Jurafsky, 2002; Punyakanok, Roth y Yih, 2008) y luego LDA (Blei, Ng y Jordan, 2013; Blei, 2012). El objetivo es extraer información semántica de los tweets que permita encontrar patrones relevantes de la información proveniente de Twitter. En específico, predecir lo que sucederá en el día d de acuerdo a los tópicos discutidos en Twitter el dia d−1. Para la evaluación consideraron los datos antes del 17 de septiembre como entrenamiento y el resto como conjunto de prueba. El modelo construido mostró un poder predictivo superior a lo que producirá un baseline considerando una predicción basada en una distribución uniforme.

Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information: En este trabajo (Wang, Browny Gerber, 2012) utilizan cuatro fuentes de datos: crímenes en Charlottesville, Virginia desde el 1ro de Marzo al 31 de Octubre del 2011, tweets escritos por la agencia de noticias CBS en el mismo periodo de tiempo. Además de información geográfica y demográfica (Wang y Brown, 2011). El objetivo principal es combinar informacion procedente de Twitter con información espacial, geográfica, temporal y demográfica para predicción de crímenes basándose en el trabajo propuesto en (Morenoff et al. 2001). Los datos geográficos contienen información como ubicaciones de carreteras, de pequeñas empresas y de escuelas. Los datos demográficos contienen información de Charlottesville asociada a bloques de censo, como población y raza. Compararon dos enfoques. El primer enfoque utiliza características numéricas que describen las propiedades geográficas y demográficas de una región. El segundo enfoque utiliza además información textual extraída de los mensajes de Twitter. El modelo híbrido propuesto fue evaluado usando datos reales de incidentes criminales para Charlottesville, Virginia. Los resultados indican que el modelo híbrido exhibe un mejor desempeño de predicción en comparación con el modelo estándar. El modelo híbrido puede generalizarse a otras áreas de aplicación donde la información textual no estructurada contiene indicadores relevantes para las propiedades espacio-temporales de los eventos. Las pruebas con datos simulados y datos reales mostraron que el algoritmo se desempeñó mejor que un modelo clásico de regresión lineal penalizada.

4.3. Data mining

4.3.1. Noticias digitales

En la literatura existen estudios que utilizan sitios web de noticias como fuente de información principal, y posteriormente aplican técnicas de text mining para clasificar eventos de interés. Estos estudios no están enfocados al ámbito de robo de automóviles, pero pueden dar luces de la metodología que se puede utilizar. En particular, hay una línea de investigación concerniente a la detección y clasificación de brotes de enfermedades infecciosas a partir de noticias web. También hay estudios relacionados con el cibercrimen. En particular, el primer estudio se detalla exhaustivamente dado que describe bien la metodología que aplican varios estudios.

Automatic online news monitoring and classification for syndromic surveillance: En este estudio (Zhang et al. 2009) se describe un sistema de monitoreo de diversas fuentes de noticias online para detectar problemas de salud pública. En particular, se desean detectar posibles brotes de enfermedades peligrosas o epidemias debido a causas naturales, errores humanos o ataques terroristas. El objetivo es diseñar y evaluar un sistema de monitoreo de noticias online y métodos de clasificación para la vigilancia en salud pública. Las fuentes de información son ProMED-mail, Argus, MiTAP y HealthMap. Estos son sistemas de noticias relacionadas con enfermedades infecciosas. Estos sistemas poseen noticias actualizadas desde medios de comunicación de todo el mundo y además, reportes internos provenientes de centros hospitalarios y ministerios de salud.

Para la adquisición de datos se utilizan técnicas de web crawling y se obtienen noticias desde 127 sitios web previamente seleccionados por expertos. Luego de recolectar las páginas HTML se procede a hacer un filtrado por palabras claves de las noticias no relacionadas. Posteriormente se aplica Text Mining, en particular utilizan las técnicas Bag of Words, Noun Phrases y Named Entities en donde cada feature es un token. Se generaron dos tipo de Features denominadas BFS (baseline feature subsets): BFS-BW (Bag of words) y BFS-Comb (combinación de las técnicas). Solo se consideran los features que aparecen más de 5 veces. Se utilizan algoritmos de machine learning para clasificar los features relevantes a partir de documentos previamente clasificados. Para la evaluación se utiliza Correlation-based Feature selection y Best First search para la generación de features. Para clasificar y evaluar se utiliza SVM y medidas habituales de accuracy, precision, recall y F-measure.

El sistema se testea con la “enfermedad de pies y boca” (FMD, por sus siglas en inglés), especialmente devastadora para animales de granjas y un peligro para la agricultura, con un evento epidémico el año 2001 en el Reino Unido que sirve para entrenar y testear. A partir de 127 fuentes de información se recolecta 1 millón de noticias directamente relacionadas con FMD. Además, se tienen 3.000 noticias recolectadas manualmente por expertos que se utilizan como gold standard. El mejor resultado del experimento indica que usando SVM a partir de los atributos con la combinación de técnicas de text mining se obtiene un accuracy de 77,04%, precisión de 77,10% y recall de 77,05%.

An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics: En este trabajo (Torii et al. 2011) se sigue la misma metodología de Zhang et al. (2009), esta vez en base a un proyecto de vigilancia biológica que cuenta con 40 especialistas que clasifican noticias manualmente. En este estudio se profundiza más en la analítica con un estudio más específico de las features usadas y de los algoritmos de machine learning. La cantidad de noticias evaluadas en el estudio es 40.000 correspondientes a fuentes de información provenientes del sur de Asia. Los resultados son similares al estudio anterior, con la ventaja de que se utilizan menos noticias.

Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web: Como se señala en el título, este trabajo (Arsevska, 2016) pretende identificar términos para detectar señales tempranas de emergencias por enfermedades infecciosas. El valor agregado de este paper es la técnica que usan para recuperar la información desde la web. Previamente entrenan modelos de text mining, obteniendo el corpus y los términos más relevantes con lo que para buscar información nueva se basan netamente en el corpus entrenado. En este caso el estudio lo realizan enfocándose en "fiebre porcina africana" con noticias entre 2011 y 2014. La metodología que aplican es la misma que en los estudios anteriores.

CybercrimeIR – A Technological Perspective to Fight Cybercrime: En este estudio (Chang, et al. 2012) las fuentes de información son sitios web de noticias y otras fuentes relacionadas con cibercrimen (bases de datos, reportes de policías, etc).Se indica que es necesario una combinación de todas las fuentes de datos posibles para lograr buenos modelos. Aplican técnicas de Information Retrieval y la metodología clásica. Se obtienen resultados por sobre el 90% de accuracy usando SVM y Naive Bayes.

Evolving fuzzy grammar for crime texts categorization: Este trabajo (Sharef y Martin, 2015) se enfoca netamente en técnicas de text mining mediante categorización de texto enfocado a incidentes criminales. El modelo de aprendizaje es construido en base a fragmentos de textos seleccionados que son luego transformados a estructuras particulares. La noción de "fuzzy grammar" se basa en que el modelo evoluciona constantemente mediante pequeños cambios incrementales. Este modelo representa un avance con respecto a los papers expuestos anteriormente en esta sección.

Real-Time News Event Extraction for Global Crisis Monitoring: En este estudio (Tanev et al. 2008) se presenta un sistema de extracción de eventos a partir de noticias online que sirva como monitoreo de crisis globales. No profundiza en teoría para la recuperación de información desde páginas web y aplica herramientas de text mining básicas. Es interesante dado que agrega características como la georeferenciación de las noticias, que no han sido abordadas por otros trabajos. Se hace una evaluación pequeña con 94 clusters de noticias en donde se encuentran 30 eventos violentos o desastrosos. El sistema fue capaz de detectar 23 de esos eventos.

En el estudio "**Online News Event Extraction for Global Crisis Surveillance"** (Piskorski, 2011) profundizan el anterior (Tanev et al. 2008) siguiendo la misma metodología, pero

concentrándose en la extracción de información, clusterización, geolocalización, selección de features y modelo de Machine Learning.

4.3.2. Visualización y Visual Analytics

En el marco de este proyecto, la investigación en visualización se enmarca principalmente en la visualización de crímenes y, más en particular, la visualización de crímenes relacionados con automóviles.

4.3.2.1. Visualización de crímenes.

Tablas y Series de Tiempo. Varias investigaciones que datan de las primeras décadas del siglo XX (1900-1940) se concentran en análisis estadísticos de distintos tipos de crímenes y la forma tradicional de visualización es través de tablas y gráficos de series de tiempo (Phelps, 1928; Phelps, 1929; Watts, 1931; Block et al. 1995; Block, 1995; Block, 1996; Lodha y Verma, 2000) . Por ejemplo, el gráfico siguiente está basado en que se muestra la evolución en cantidad de distintos crímenes entre 1897 y 1927 en Rhode Island, EEUU. El mismo autor presenta en (Phelps, 1929) un análisis más detallado de distintos crímenes (contra personas, contra propiedad, contra moralidad del sexo, etc.) comparándolo con evolución de índices de pobreza, usando nuevamente series de tiempo. En (Watts, 1931) lleva a cabo un análisis similar, esta vez comparando la evolución de crímenes contra densidad poblacional.

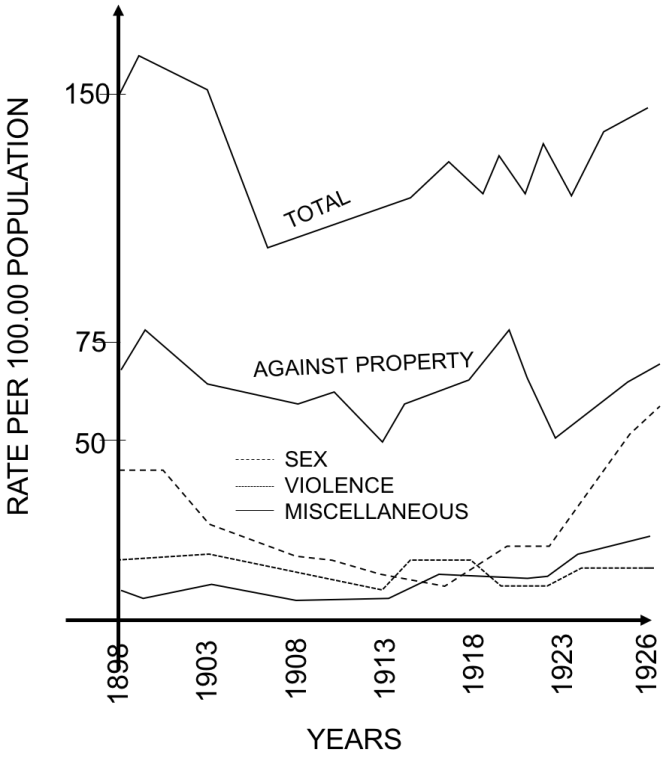

Figura 1. Frecuencia de crímenes en Rodhe Island 1897-1927, adaptado de Phelps (1928)

Mapeo Espacial: Una de las formas más comunes para el análisis visual de crímenes corresponde a la agregación de datos y posterior visualización espacial de esta información en mapas, lo que en inglés se conoce como *crime mapping*. Si bien estos mapas se pueden encontrar desde hace casi 3 décadas, es en los años '90, con el auge de las tecnologías

GIS (Geographic Information Systems), donde se popularizaron en el dominio de crime mapping (Block 1995, Block et al. 1995, Block, 1996). En base a este concepto, Lodha y Verma (2000) presentaron un avanzado sistema de análisis de crimen con varios componentes de visualización que mapeaba datos de incidentes sobre una ciudad usando (i) gráficos de barras apiladas, y (i) mapas con información multicapa para comparar patrones de día y noche. Otros trabajos interesantes de mencionar con mapeos espaciales son Cameron (2001), Ratcliffe (2004), Hagenauer et al. (2011). Cameron (2001) realiza un análisis estadístico y visualización espacial de crímenes enfocándose en Appalachia. Rathcliffe (2004) revisa las tecnologías de crime mapping a mediados de la década de 2000s y propone que las fuerzas de orden se capaciten en tecnologías relacionadas como GIS. En Hagenauer et al. (2011) estudian el impacto de un huracán en la variación de crimen, así como las trayectorias de los crímenes usando Self-Organizing Maps (SOM).

Visualización de redes: Este tipo de diagramas (grafos, también llamados "diagramas nodo-enlace") se ha usado principalmente para estudiar la estructura de organizaciones criminales, obteniendo grupos, analizando centralidad de nodos y evolución de comunidades (Xu y Chen, 2005). Uno de los sistemas más citados es COPLINK (2003), que presentaba en una estructura de red personas involucradas en actividades delictivas y sus relaciones. Complementando este trabajo, Chen et al (2005) combina información espacial y de red que permitan buscar y organizar la información de crímenes desde distintas vistas de visualización.

Portales Open Data: Ciertos gobiernos municipales, como las ciudades de Chicago, Washington D.C. y Pittsburgh han implementado portales que permiten de forma pública y abierta acceder a información y visualizarla. Sus portales pueden servir de inspiración para algunas de las visualizaciones y herramientas de analítica visual de este proyecto, como el portal de la ciudad de Pittsburgh, PA (https://tools.wprdc.org/) o el de la ciudad de Chicago. (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2)

4.3.2.2. Visualización de crímenes de automóviles

Con respecto al uso de visualización para análisis de crímenes relacionados con vehículos, los trabajos hacen uso de las técnicas y herramientas usadas en visualizar y analizar crimen en general, pero algunos se enfocan específicamente en este problema son. Kursun et al. (2005) presenta un sistema que combina un modelo clustering con una visualización GIS para agrupar los distintos tipos de crímenes relacionados con robo de vehículo, pero no presentan una evaluación de su modelo, más bien se enfocan en presentar la implementación de un sistema. Otro trabajo, esta vez de Lu et al. (2003) estudia un dataset de vehículos robados en Buffalo, NY, que contiene información de las rutas y distancias de los vehículos. El análisis muestra que mientras mayor es la distancia recorrida por el delincuente, menor es la probabilidad de recuperarlo, cayendo de forma exponencial. La distancia media recorrida de vehículos recuperados era de alrededor de 3,7 km. Más recientemente, Wang et al. (2012) presentan una herramienta para recuperación de información y minería de datos STT (Spatio-Temporal Textual). Esta herramienta combina información de dos datasets (Crimen en Washington, DC area del 2006 al 2010 y un dataset de terrorismo global del 2004 al 2010) y usando modelos de recuperación de información,

topic modeling, y visualización usando nubes de palabras y mapas, presenta tendencias de diferentes crímenes, incluyendo robo de autos y robo de especies desde vehículos.

5. ROBO DE AUTOS Y ACCESORIOS EN CHILE

En este capítulo se presenta la caracterización preliminar y un análisis de patrones de los datos de robos de vehículos y accesorios. La sección 5.1 analiza las fuentes de AACH-DERV, Twitter y medios noticiosos en línea. Luego en la sección 5.2 se presenta un análisis exploratorio en diferentes dimensiones.

5.1. Análisis de fuentes de datos

5.1.1. Registro de robo de vehículos asegurados (AACH-DERV)

La base de datos de la Asociación de Aseguradoras de Chile cuenta con el registro de vehículos que estando asegurados en algunas de las compañías de seguros en Chile han sido sustraídos, hurtados o robados[1] entre los años 2011 y lo que transcurre del año 2017. No obstante, todos los análisis que se presentan en este documento incorpora sólo los datos comprendidos entre los años 2011 y 2016, esto ya que permite hacer una comparación exacta entre los periodos de tiempo comprometido en cada año. Los datos con los que se realizan los análisis suman 55.626 registros distribuidos según el año como se muestra en la siguiente figura.

[1] De aquí en adelante nos referiremos a la sustracción, hurto o robo meramente como robo. Esto sin ánimo de imponer un carácter violento a todos los delitos, sino más bien por simplicidad.

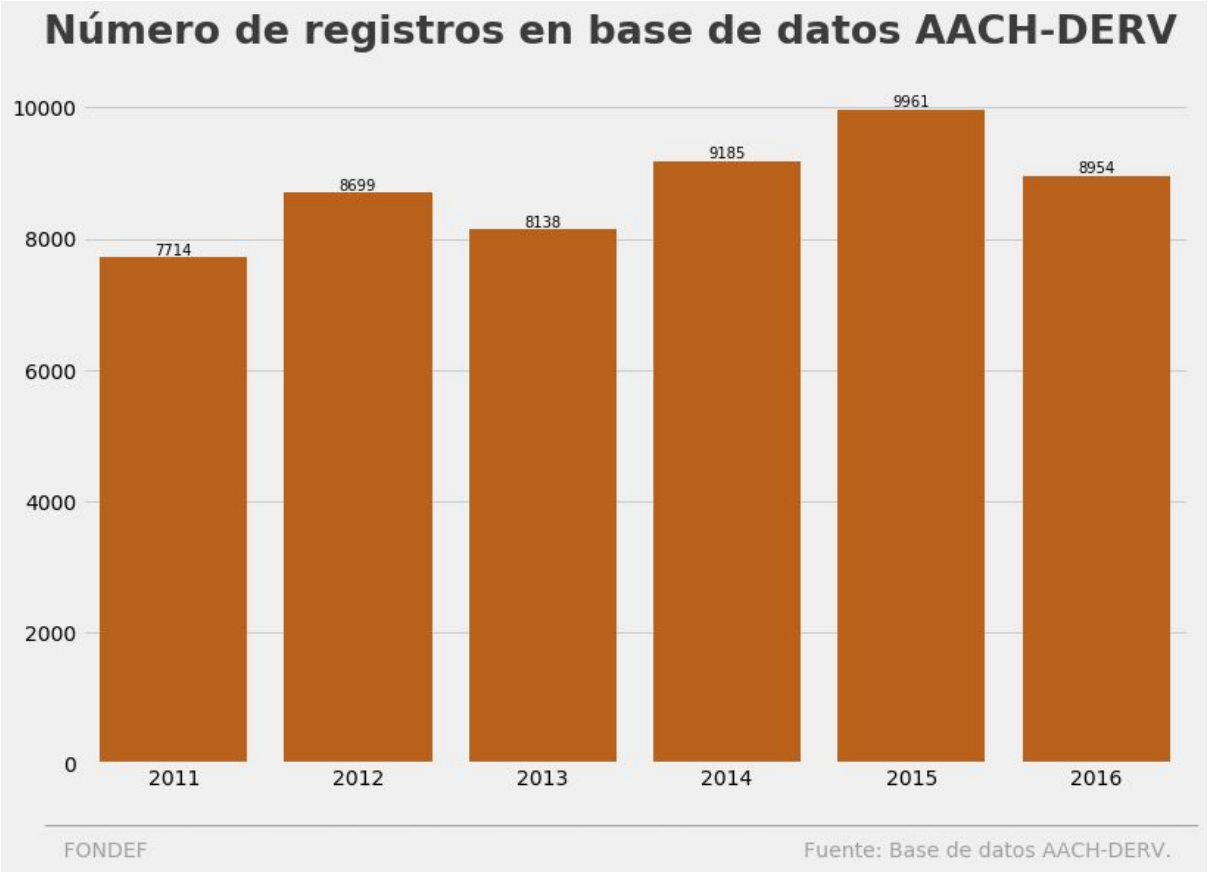

Figura 2. Cantidad de robos registrados por año en base de datos AACH-DERV

5.1.1.1. Detalles Base de Datos AACH-DERV

La base de datos cuenta con 74 campos distintos, cada uno de estos aporta información que caracteriza al vehículo, el lugar donde fue robado, la víctima, la fecha y hora del robo, además del relato de la víctima sobre el delito. No obstante, de estos 74 campos, algunos son información redundante que codifica o re codifica otro campo, otros son información confidencial y sólo una pequeña proporción será de interés para los análisis que se pretenden emplear en esta sección. La siguiente tabla explica la información contenida en los campos.

Nombre	Tipo de dato	Filas no null	Descripción
id_prose	ordinal	55626	ID única transacción.
sin_patente	nominal	55626	Patente vehículo siniestrado.
sin_fecha_siniestro	fecha	55626	Fecha (día, mes y año) en donde ocurrió el siniestro de robo.
ase_abreviatura	nominal	55626	Abreviatura de aseguradora en la cual se registra el robo.
obs_ult_estado	nominal	3179	Validación interna.

sin_siniestro	ordinal	55626	Número de siniestro interno de la compañía. Cada número es independiente.
sin_fecha_denuncia	fecha	55626	Fecha denuncia siniestro.
sin_lugar_siniestro	ordinal	55626	ID Tipo de lugar en donde ocurrió el siniestro.
lug_desc	nominal	55626	Tipo de lugar en donde ocurrió el siniestro.
rpc_id_region_siniestro	ordinal	55626	ID Región en donde ocurrió el siniestro.
reg_descripcion_region	nominal	55624	Región en donde ocurrió el siniestro.
rpc_id_provincia_siniestro	ordinal	55626	ID Provincia en donde ocurrió el siniestro
pro_descripcion_provincia	nominal	55624	Provincia en donde ocurrió el siniestro
rpc_id_comuna_siniestro	ordinal	55626	ID Comuna en donde ocurrió el siniestro
reg_descripcion_comuna	nominal	55624	Comuna en donde ocurrió el siniestro
sin_direccion_siniestro	nominal	46819	Relato de dirección del siniestro.
sin_hora_siniestro	hora	55626	Hora de ocurrencia del siniestro.
est_id_prose	ordinal	55626	Validación interna.
esp_desc	nominal	55626	Validación interna.
sin_fecha_ultimo_estado	fecha	55626	Validación interna.
sin_poliza	nominal	50793	Número de poliza.
sin_item	ordinal	48875	Validación interna.
sin_check_registro_civil	nominal	55626	Validación interna.
sin_concurrencia	nominal	55626	Validación interna.
sin_ind_caso_raro	nominal	55626	Validación interna.
sin_glosa_caso_raro	nominal	2	Validación interna.
tve_id_tipo_vehiculo	ordinal	55626	ID Tipo de vehículo según registro civil
tve_desc	nominal	55626	Tipo de vehículo según registro civil
mar_id_vehiculo	ordinal	55626	Identificación interna de la marca del vehículo siniestrado.
mar_desc	nominal	55626	Marca del vehículo involucrado en el robo.
mod_id_vechiculo	ordinal	55626	ID del modelo asociado al vehículo siniestrado.
mod_desc	nominal	55626	Modelo del vehículo siniestrado.
sin_ano_vechiculo	ordinal	55626	Año de fabricación otorgado por el registro civil.
sin_motor_vehiculo	nominal	53456	Motor asociado al vehículo robado.

sin_chasis_vehiculo	nominal	53200	Chasis del vehículo asegurado
cve_id_vehiculo	ordinal	55626	ID del color del vehículo.
cve_desc	nominal	55626	Color del vehículo asegurado.
sin_valor_comercial_veh	ordinal	55626	Valor comercial del vehículo, sin detalle de las unidades utilizadas.
sin_dispositivos_vehiculo	nominal	55626	Validación interna.
sin_dispositivos_desc	nominal	188	Validación interna.
rpc_id_region_asegurado	ordinal	55626	ID Región en donde ocurrió el siniestro.
reg_descripcion_region_asegurado	nominal	55626	Región de residencia de la persona que contrató el seguro.
rpc_id_provincia_asegurado	ordinal	55626	ID Comuna de residencia de la persona que contrató el seguro.
pro_descripcion_provincia_asegurado	nominal	55626	Comuna de residencia de la persona que contrató el seguro.
rpc_id_comuna_asegurado	ordinal	55626	ID Comuna de residencia de la persona que contrató el seguro.
reg_descripcion_comuna_asegurado	nominal	55626	Comuna de residencia de la persona que contrató el seguro.
rpc_id_region_conductor	ordinal	55626	ID Región de residencia de la persona que conducía el auto al momento del siniestro.
reg_descripcion_region_conductor	nominal	55626	Región de residencia de la persona que conducía el auto al momento del siniestro.
rpc_id_provincia_conductor	ordinal	55626	ID Provincia de residencia de la persona que conducía el auto al momento del siniestro.
pro_descripcion_provincia_conductor	nominal	55626	Provincia de residencia de la persona que conducía el auto al momento del siniestro.
rpc_id_comuna_conducto	ordinal	55626	ID Comuna de residencia de la persona que conducía el auto al momento del siniestro.
reg_descripcion_comuna_conductor	nominal	55626	Comuna de residencia de la persona que conducía el auto al momento del siniestro.
sin_comisaria	nominal	45121	Relato de comisaria en donde se constató el heho.
sin_fecha_parte	nominal	55626	Fecha del parte registrado en comisaria.
sin_tribunal_fiscalia	nominal	3061	Tribunal y/o fiscalía que adscribe al caso siniestrado (recibido por carabineros).
sin_fecha_citacion	nominal	55626	Validación interna.
sin_relato	nominal	51451	Relato del siniestro.

car_id	ordinal	55626	Registro del vehículo asegurado reportado.
cre_id	ordinal	55626	Validación interna.
sin_activo	nominal	55626	Validación interna.
sin_cargado_bizagi	nominal	55626	Validación interna.
id_PROSE_BPM_ROBO	ordinal	55626	Validación interna.
UltimoOrden	nominal	49258	Validación interna.
FechaOrden	nominal	49258	Validación interna.
UsuarioBPM	nominal	627	Validación interna.
CargadoBPM	nominal	55626	Validación interna.
PrimeraFechaValidacionProse	fecha	55624	Validación interna.
EncontradoOtraVia	nominal	55626	Informa si la información del hallazgo fue distinta a carabineros.
tareasBizagi	nominal	48791	Validación interna.
FechaCarga	nominal	55626	Fecha en que se agregó al sistema (de la aseguradora) el robo vehicular

Tabla 1. Resumen y descripción de los campos utilizados en los análisis

5.1.1.2. Análisis en el tiempo y espacial (por región)

En la Figura 3 se caracteriza el monto del vehículo sustraído, donde es posible apreciar que existe una alta concentración en torno a los 7.000.000 de pesos. No obstante, más del 75% de los valores observados tenían asignado un valor de 0 pesos, por lo cual no se puede descartar la opción de que exista un sesgo de selección en el campo que registra el valor de los vehículos robados.

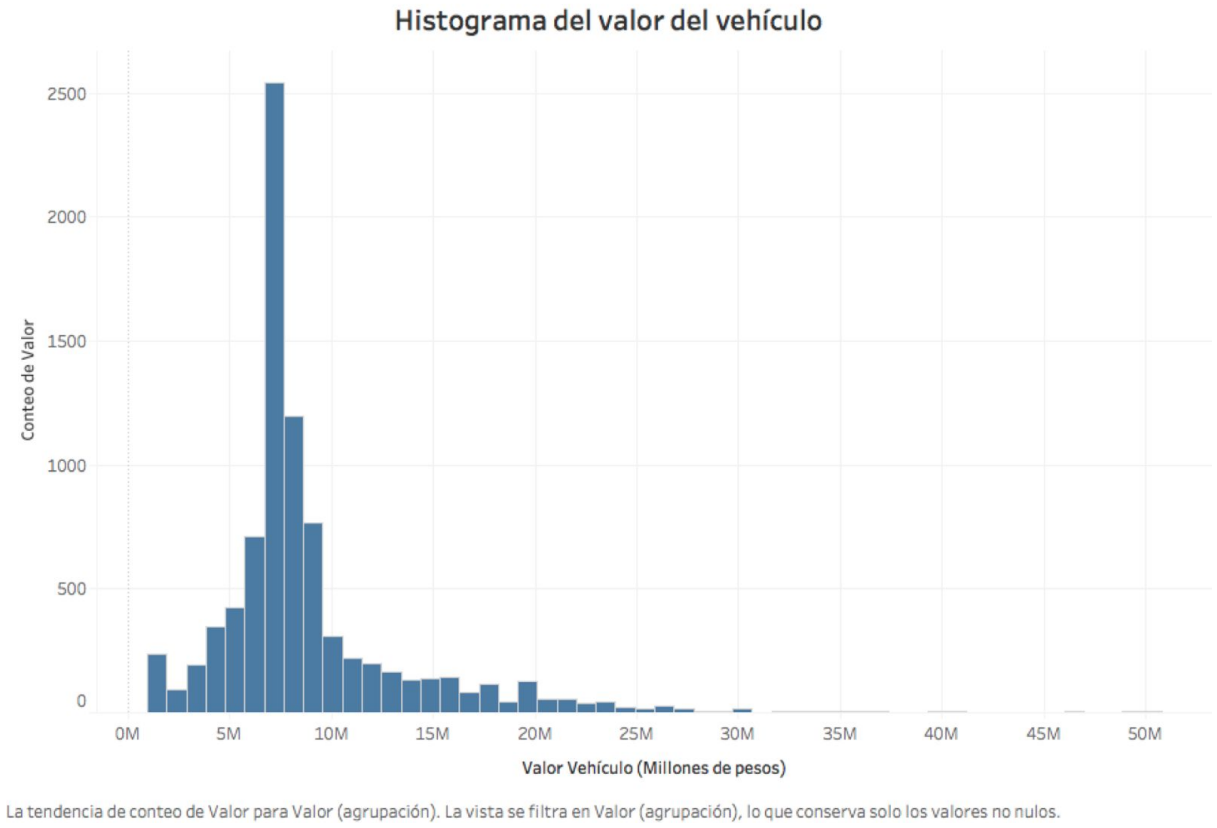

Figura 3. Distribución del monto del vehículo sustraído

En la Figura 4 se analiza el tiempo entre la fecha en las que se constata que ocurrió el delito y la fecha en la que se denunció este delito. En la base de datos se registran valores sumamente extremos, pero se analiza hasta 31 días de diferencia entre estos sucesos. Se estima que en promedio transcurre un día entre que sucede el delito y se denuncia en las autoridades pertinentes.

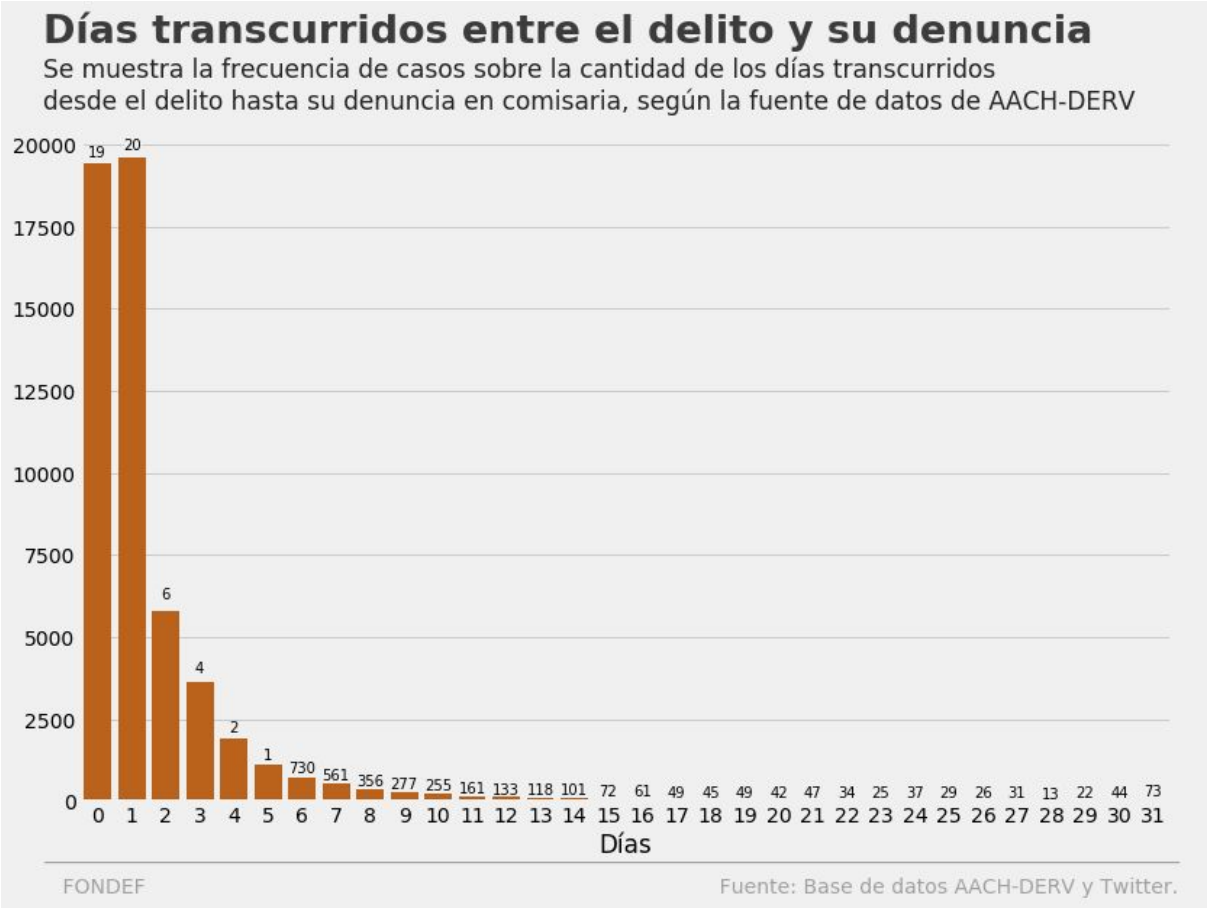

Figura 4. Tiempo entre el delito y su denuncia

La Figura 5 presenta un resumen de los robos de vehículos que se observan en la base de datos de AACH-DERV según la región en la que esta se registra el delito.

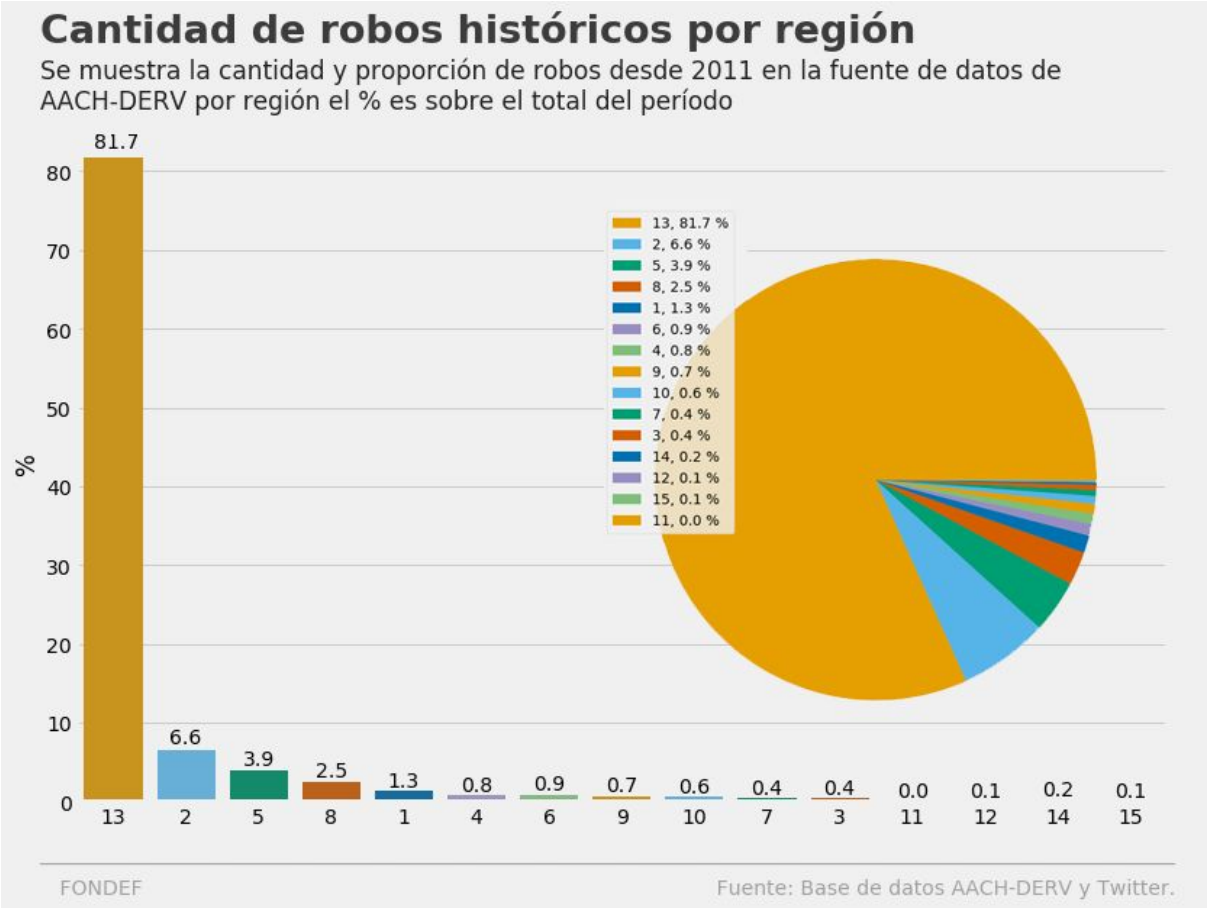

Figura 5. Distribución del robo de vehículos por región

En la Figura 6 el **81,7% de los robos vehiculares son registrados en la Región Metropolitana**, lo cual en teoría puede ser consistente con el hecho de que cerca del **60% del parque automotriz se presenta en dicha región**, además del hecho de que las tasas de criminalidad suelen experimentar una proporcionalidad superior a la población registrada en un lugar (Danzinger, 1976; Morenoff et al. 2001). No obstante, es necesario comprender además que esta base de datos representa los vehículos asegurados y puede existir un sesgo en el parque automotriz de vehículos asegurados en la Región Metropolitana comparado con el resto de las regiones.

La Figura 7 muestra la evolución de la distribución del robo de vehículos por región como porcentaje del total de robos registrados en el año. De este gráfico se puede concluir que los robos registrados a lo largo de Chile ha mantenido una proporción constante desde el 2012 al 2016. Por tanto, en el periodo comprendido entre estos años se descarta la existencia de una estacionalidad o efecto exógeno sobre alguna de las regiones. Más bien, los robos registrados corresponden a un efecto sistemático.

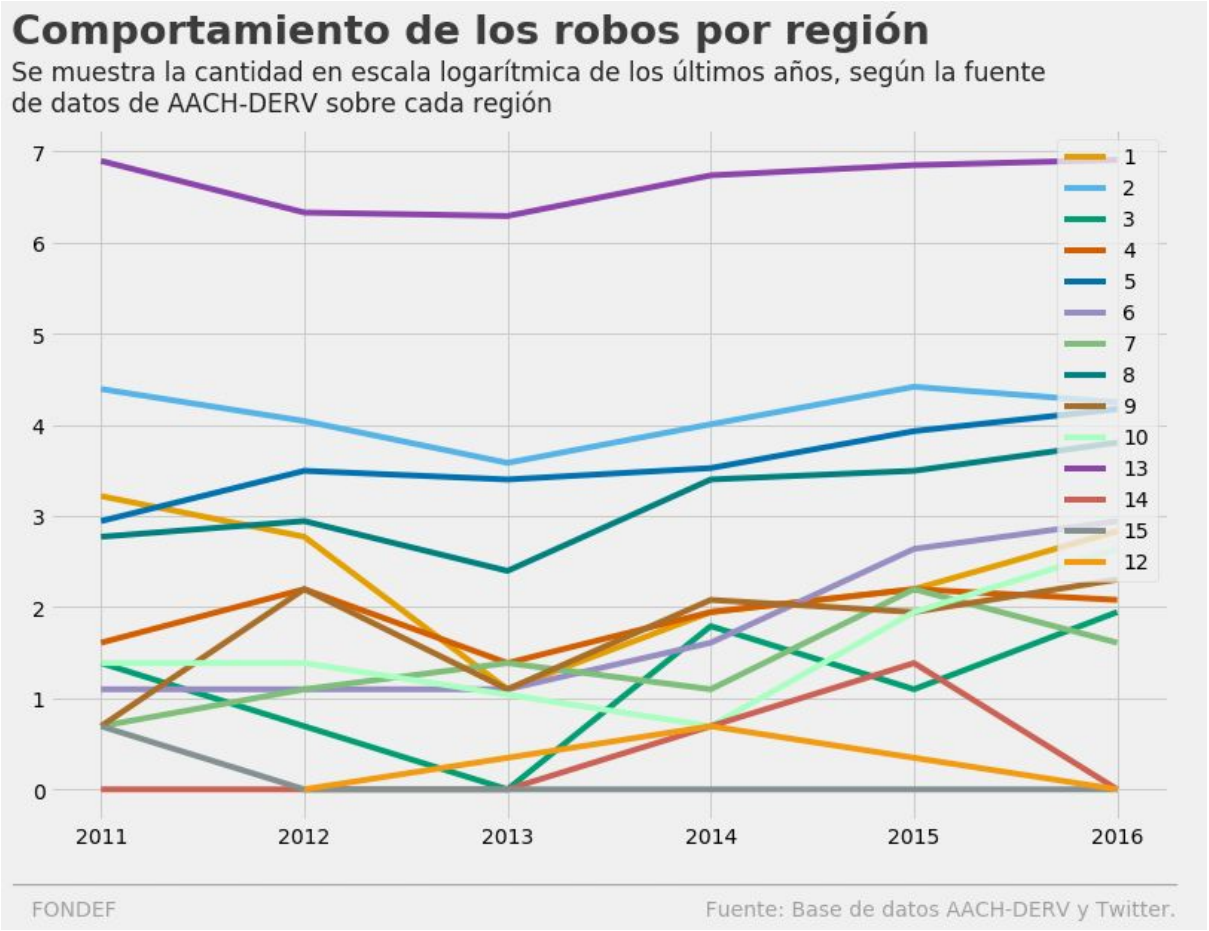

Figura 7. Evolución de la distribución del robo de vehículos por región como porcentaje

5.1.1.3. Análisis de vocabulario en denuncias

Parte importante de las denuncias es el relato de la víctima. A continuación se presenta una nube de las palabras más frecuentes en los relatos y sus respectivas frecuencias.

Figura 8. Nube de palabras de los relatos prestados por las víctimas

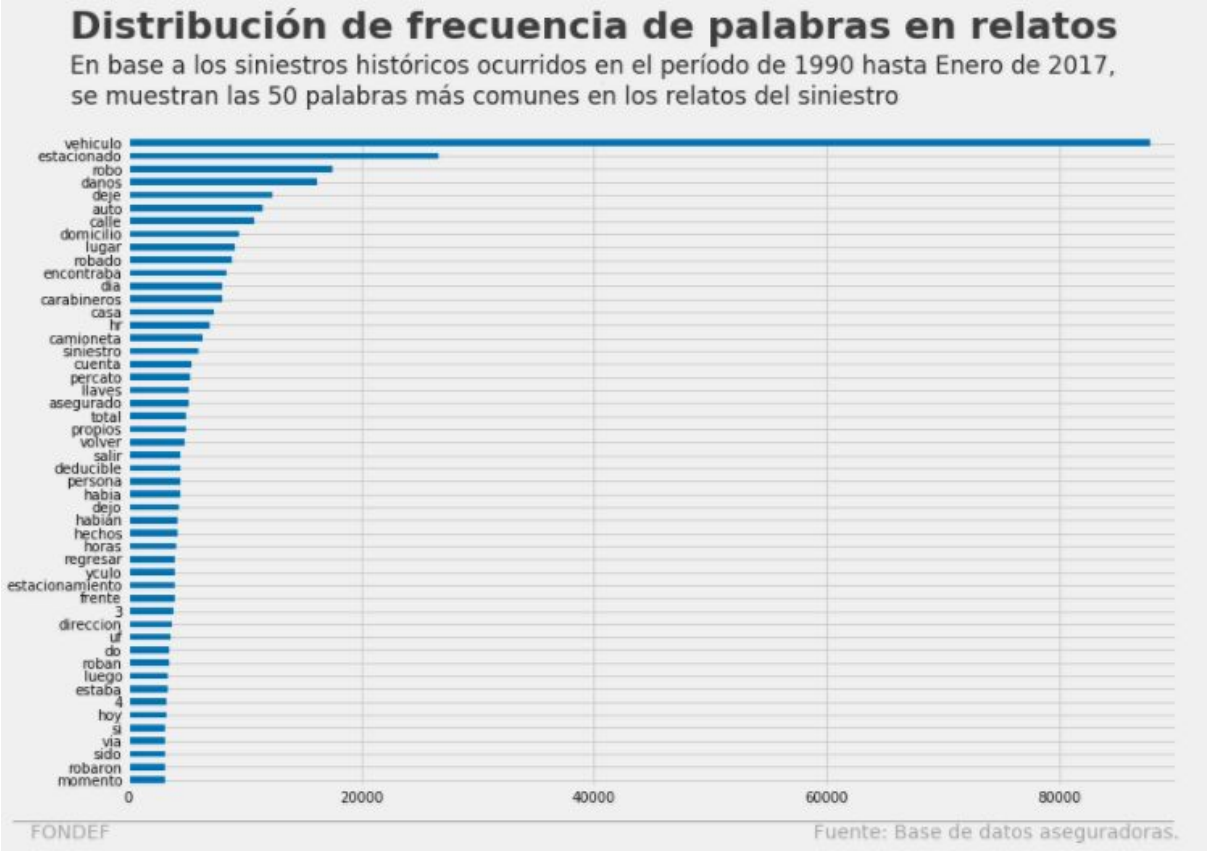

Figura 9. Frecuencia de las palabras en los relatos de las víctimas

5.1.2. Twitter

El objetivo es explorar qué tipo de información se puede extraer desde fuentes de datos no-oficiales que son generadas de forma autónoma por las personas. En particular nos centramos en medir la relación entre los datos encontrados en la red social Twitter con respecto a los datos de las compañías de seguros.

5.1.2.1. Recolección de datos

Para recolectar datos de Twitter se creó un script en el lenguaje de programación R que realiza consultas de búsqueda en la página de Twitter con la siguiente consulta:

```
query[(robar | robado | robaron | robo | robada) & patente]
```

Con esta consulta se obtienen los mensajes publicados por los usuarios que coinciden con los la consulta especificada. Para cada uno de estos mensajes se extrajo su número de patente (si es que esta había sido incluída en el mensaje del usuario). Con esta información se consultó el sitio: http://www.multidata.cl/warpit/html/main/main/. Este sitio entrega como respuesta los siguientes parámetros:

A. Tipo de vehículo (AUTOMÓVIL, STATION WAGON, CAMIONETA, TODO TERRENO, FURGON),

B. Marca
C. Modelo
D. Año
E. Color

De forma adicional, se obtuvo información asociada a la tasación del vehículo desde el sitio de Impuestos internos http://www.sii.cl/pagina/actualizada/noticias/tv_historica.htm
A estos datos fueron preprocesados aplicando técnicas de limpieza estandarización de texto y eliminación de duplicados, resultando en 7.267 registros.

5.1.2.2. Análisis de los datos

En las Figuras 10 y 11 se muestra el solapamiento que existe entre los datos obtenidos desde Twitter y los datos oficiales de las aseguradoras. De los 39.745 casos reportados en la AACH-DERV 1.763 también fueron reportados en Twitter. De los 7.267 casos reportados en Twitter 5.504 fueron casos que no aparecen en AACH-DERV. Tipos de vehículos más reportados en ambas fuentes de datos. Se puede apreciar distribución muy similar entre los tipos de vehículos que se reportan.

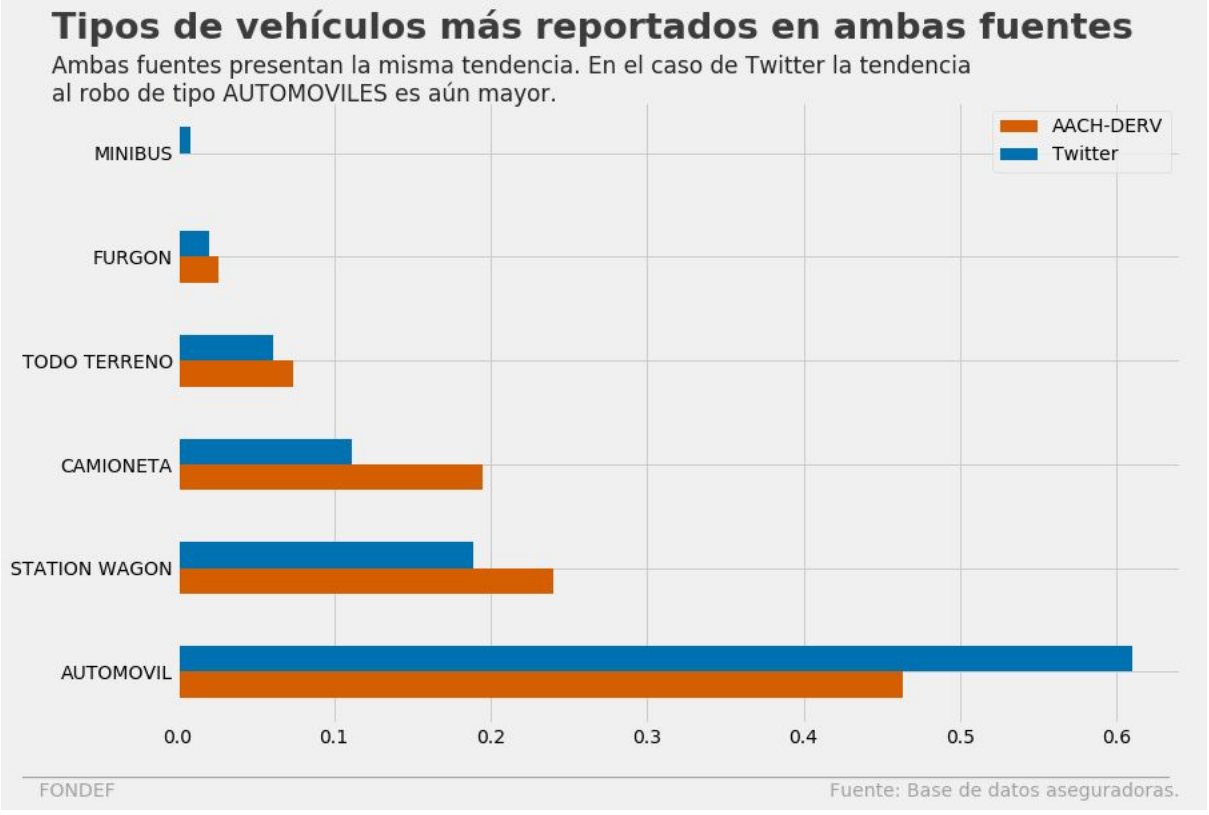

Figura 10. Reporte de robos más frecuentes en Twitter y AACH-DERV

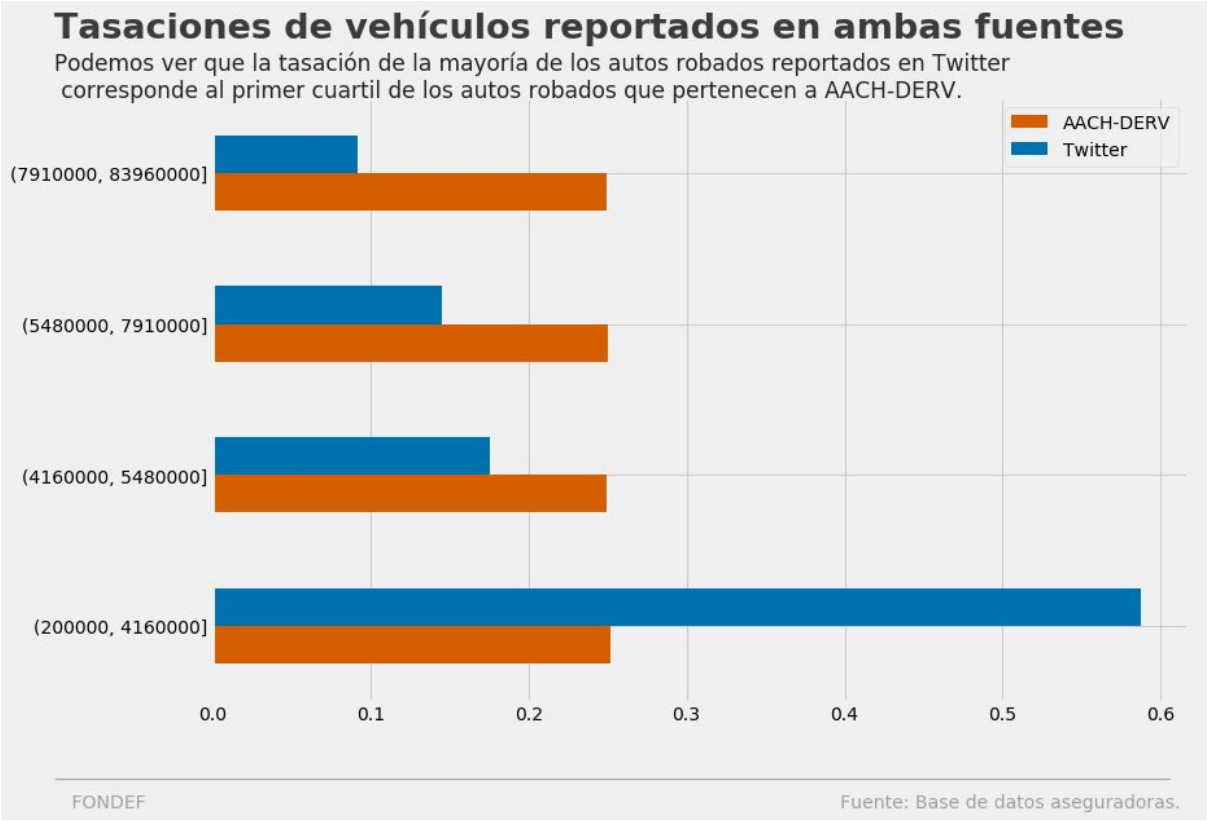

Figura 11. Relación con la tasación de los vehículos en ambas fuentes de datos

El gráfico anterior demuestra que la mayoría de los autos robados reportados en Twitter tiene una tasación inferior a los autos reportados en AACH-DERV. Lo cual parece intuitivo, ya que Twitter puede contener reportes sobre autos que están en el parque automotriz completo, sin embargo AACH-DERV solo incluye un segmento de autos asegurados. En general dentro de los autos asegurados están los vehículos de mayor valor del parque automotriz.

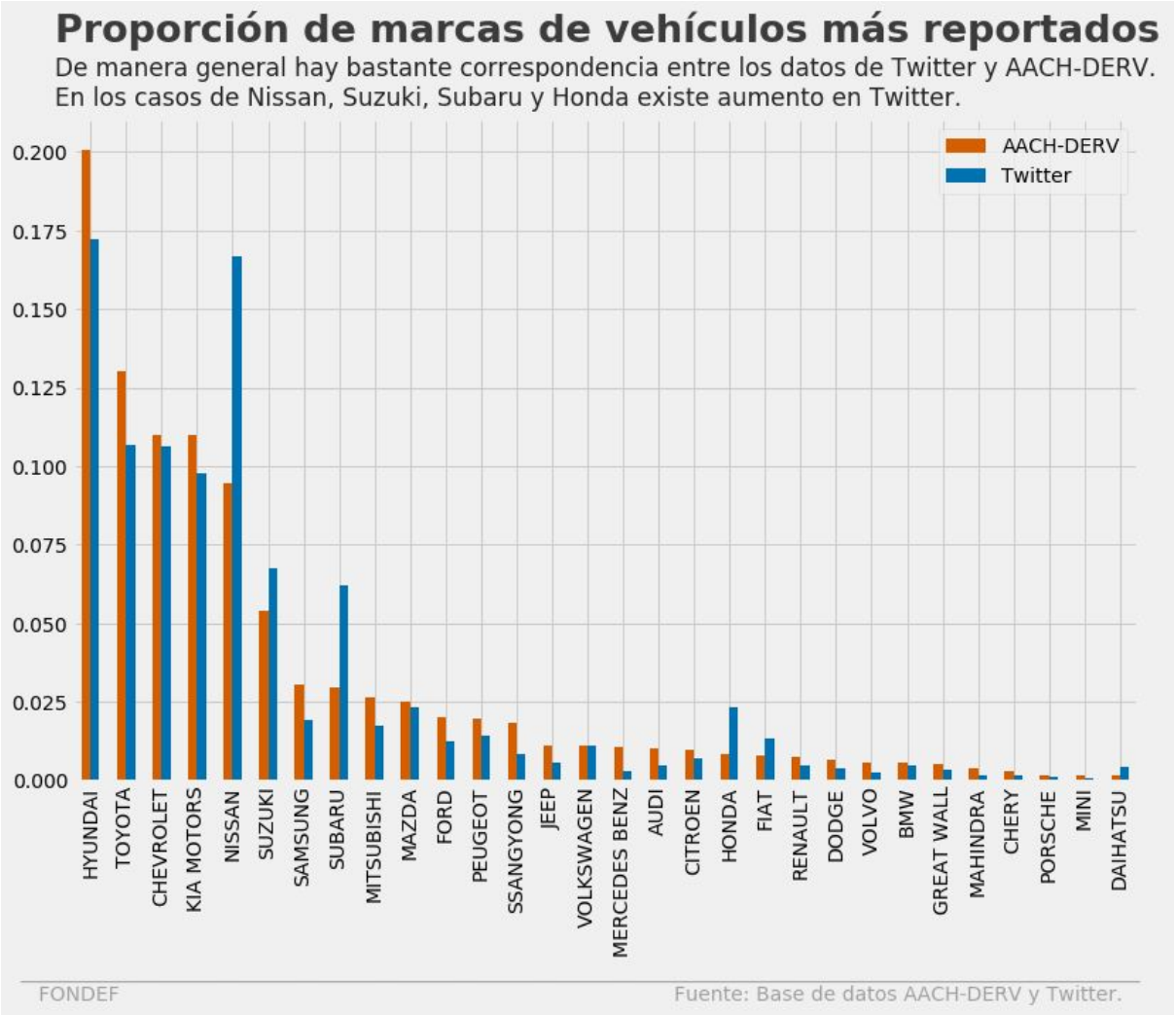

Figura 12. Marcas de autos más frecuentes en los reportes en ambas fuentes de datos

De manera general hay bastante correspondencia entre ambas fuentes de datos, con una correlación de 0,74. En los casos de NISSAN, SUZUKI, SUBARU y HONDA podemos ver una tendencia a un aumento en los reportes en Twitter.

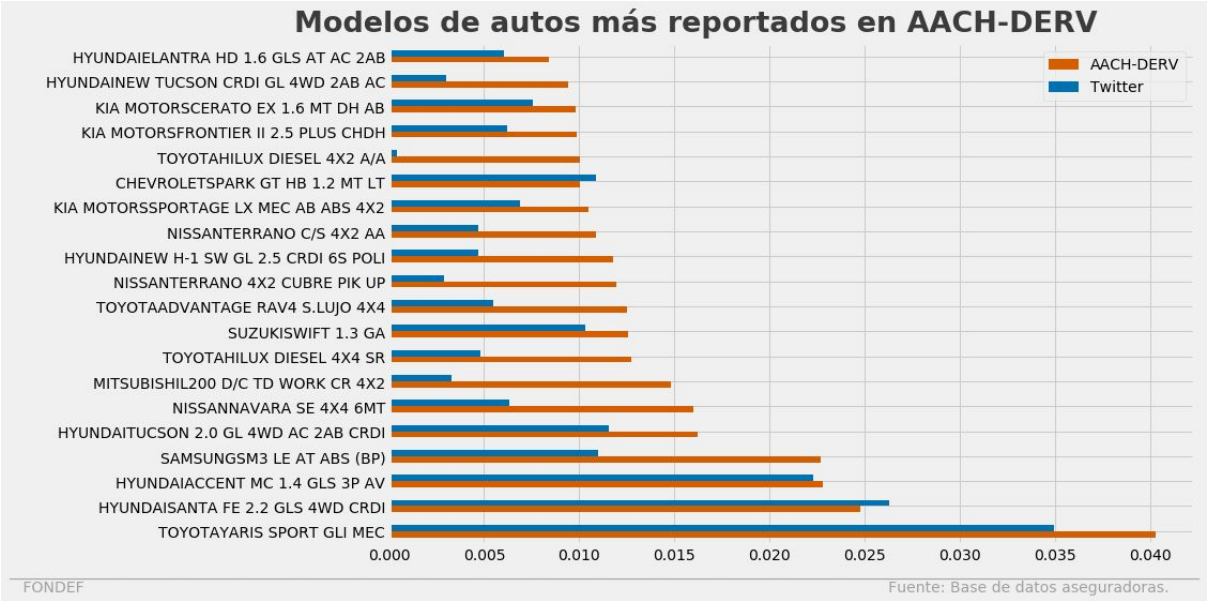

Figura 13. Modelos de autos más frecuentes en los reportes de AACH-DERV

Existe bastante correspondencia entre ambas fuentes de datos, con una correlación de 0,74 de manera general mientras que para los 20 más comunes en AACH-DERV que se muestran en el gráfico anterior la correlación es de 0,89. En los casos de NISSAN, SUZUKI, SUBARU y NISSAN podemos ver una tendencia a un aumento en los reportes en Twitter.

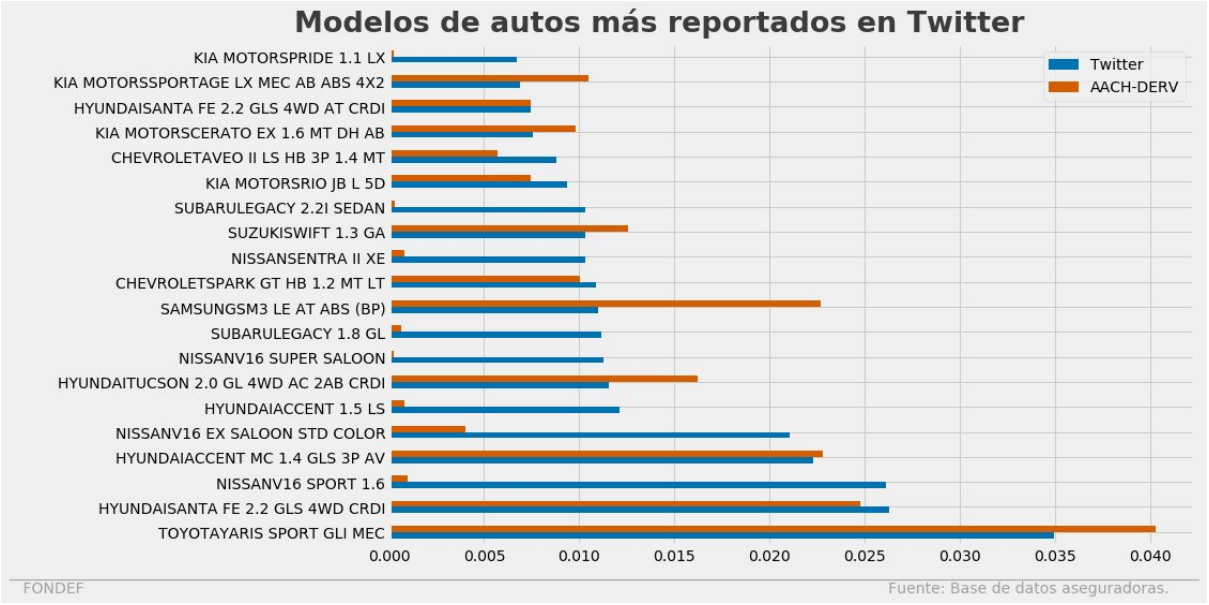

Figura 14. Modelos de autos más frecuentes en los reportes de Twitter

En el caso de los 20 modelos más reportados en Twitter la correlación es de 0,60. Los modelos NISSAN V16, HYUNDAI ACCENT y SUBARU LEGACY muestran una clara tendencia a aparecer en reportados mucho más en Twitter que en la AACH-DERV. Vehículos que en su mayoría son utilizados como taxis, y no están asegurados.

Análisis de patrones temporales

La movilidad humana está asociada a patrones temporales en diferentes granularidades, como día de la semana y hora del día. En la mañana los niños son llevados a la escuela y el movimiento de personas se produce desde la casa hacia el lugar de trabajo. Los fines de semana se realizan actividades en familia, mientras que en las tardes y las noches se realizan actividades más relacionadas al esparcimiento y la vida nocturna de la ciudad. Este tipo de actividades pueden estar asociadas a patrones de robos de autos. Veamos a continuación un análisis de patrones temporales en ambas fuentes de datos donde el número de robos por mes es normalizado de acuerdo al total desde al 2012 al 2016 en ambas fuentes de datos.

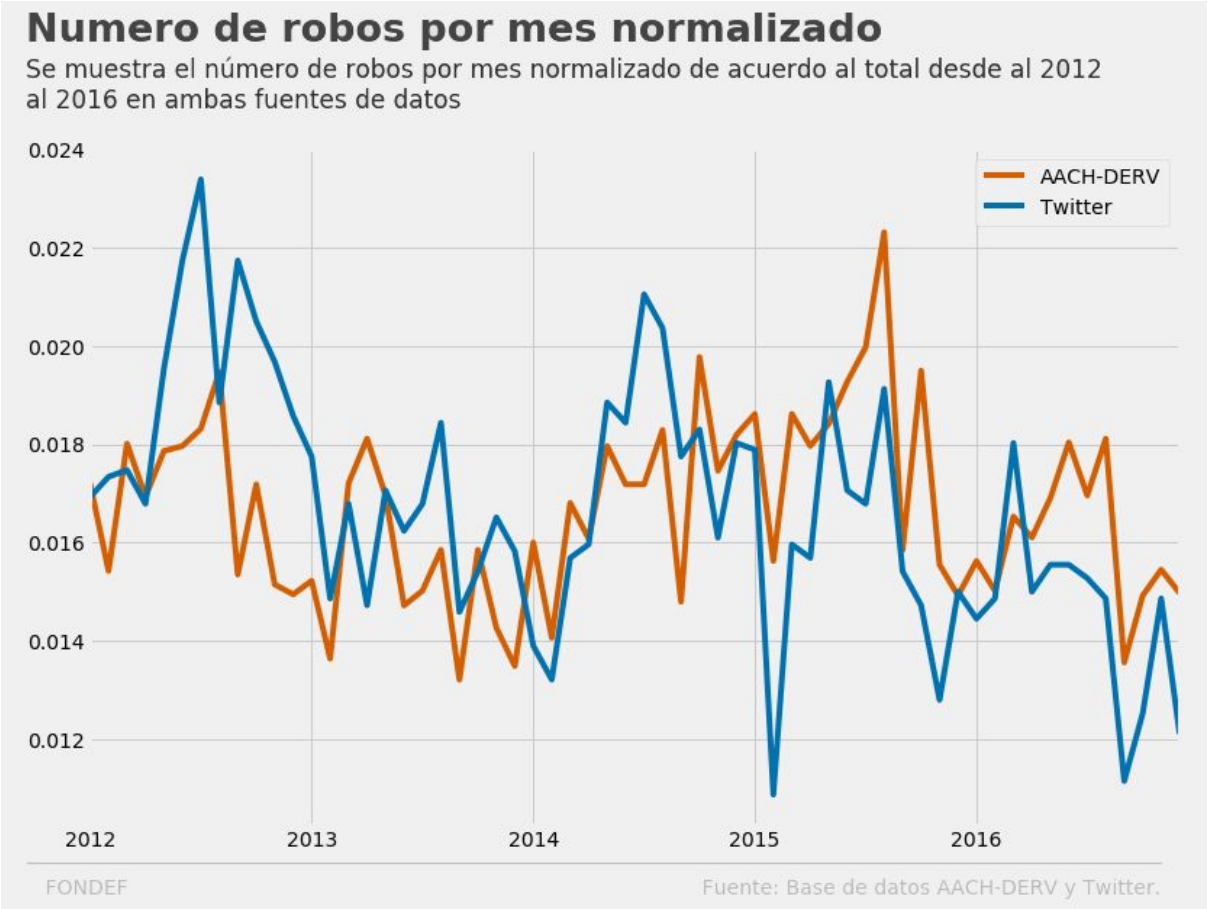

Figura 15. Número de robos mensuales en ambas fuentes de datos

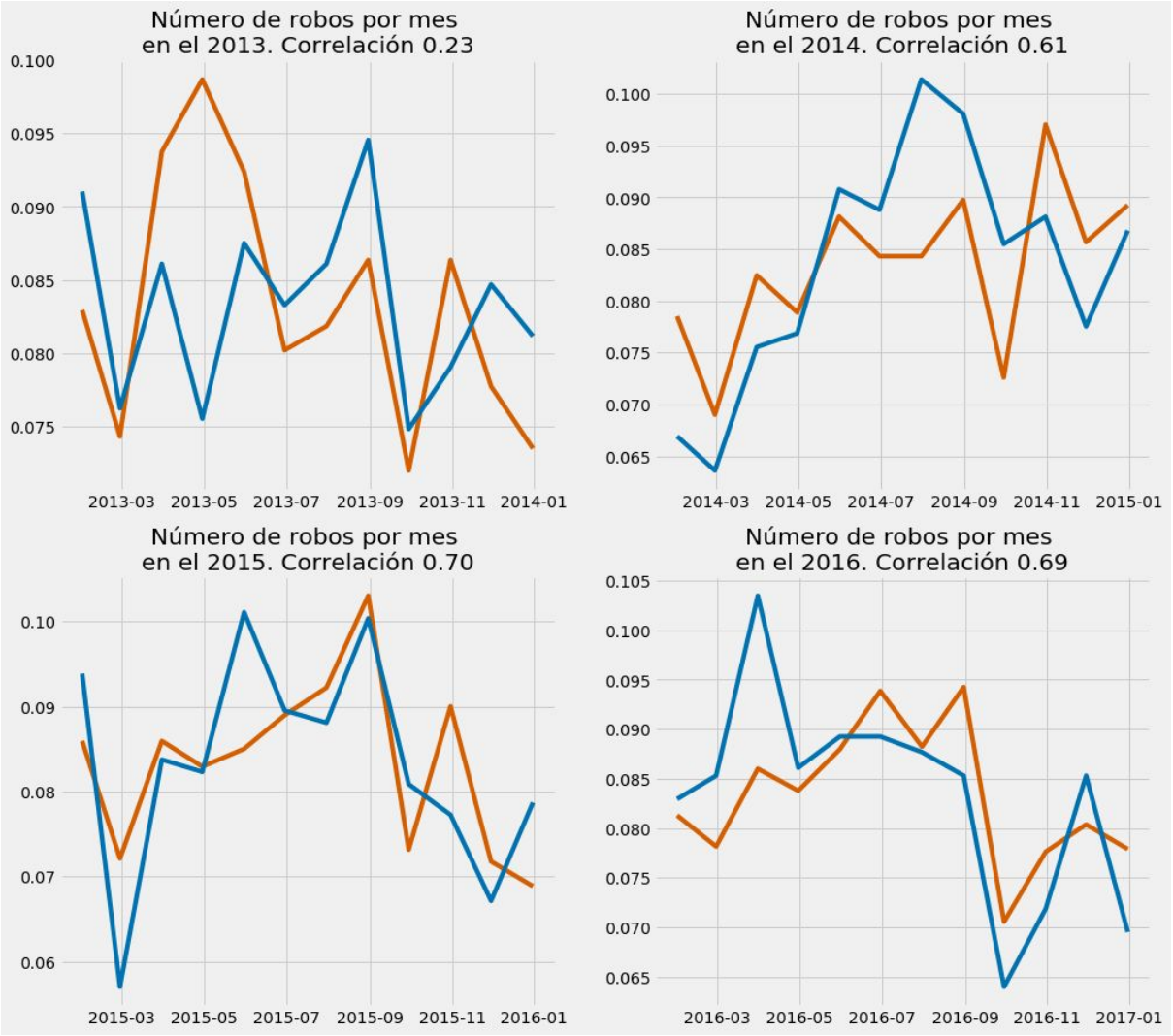

Figura 16. Número de robos mensuales por año en ambas fuentes de datos

Al separar los datos por años podemos observar un aumento en la correlación entre el volúmen de robos reportados en Twitter y AACH-DERV, llevando hasta un nivel aproximado de 0.7 de correlación en el tiempo. Esto indica que Twitter, a pesar de ser una fuente no estructurada y muy ruidosa, permite hacer una estimación del volúmen de robos que son reportados en las aseguradoras.

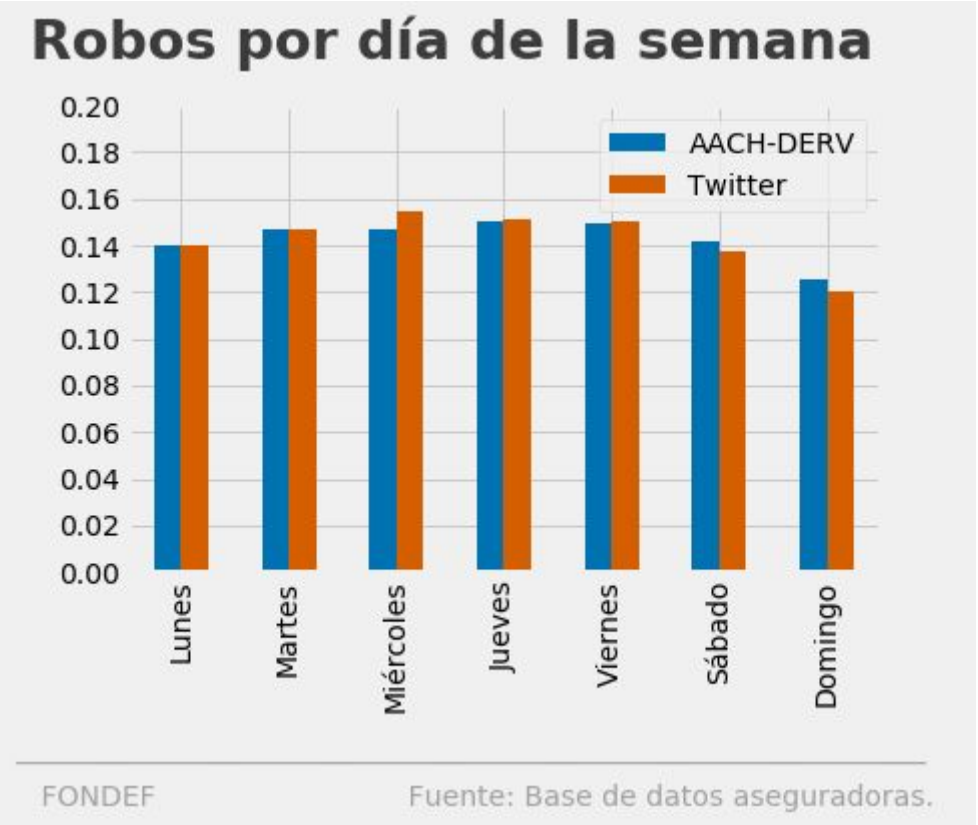

Figura 17. Registro de robos por día de la semana en ambas fuentes de datos

Una relación similar se da en cuanto a la distribución de robos por días de la semana. El gráfico anterior muestra los robos por día de la semana de todos los reportes en ambas fuentes de datos, en el caso de Twitter el momento en que se publica. En general, los robos aumentan hasta mediados de semana, disminuyendo gradualmente hacia los fines de semana.

Los siguientes gráficos tienen como objetivo analizar cómo se comportan los robos por dia de la semana y por hora del dia. La información asociada a la hora en ambas fuentes de datos no es un indicador real del momento en que ocurrió el robo. En el caso de Twitter, es el momento en que se registra el tweet, mientras que en el caso de AACH-DERV, si bien es la hora del momento que reporta el dueño del auto, es de esperar que sea un valor aproximado, pues en muchos casos no existe constancia real del momento exacto sino una estimación del momento en que ocurrió el robo.

Figura 18. Robos por hora del día y día de la semana en AACH-DERV

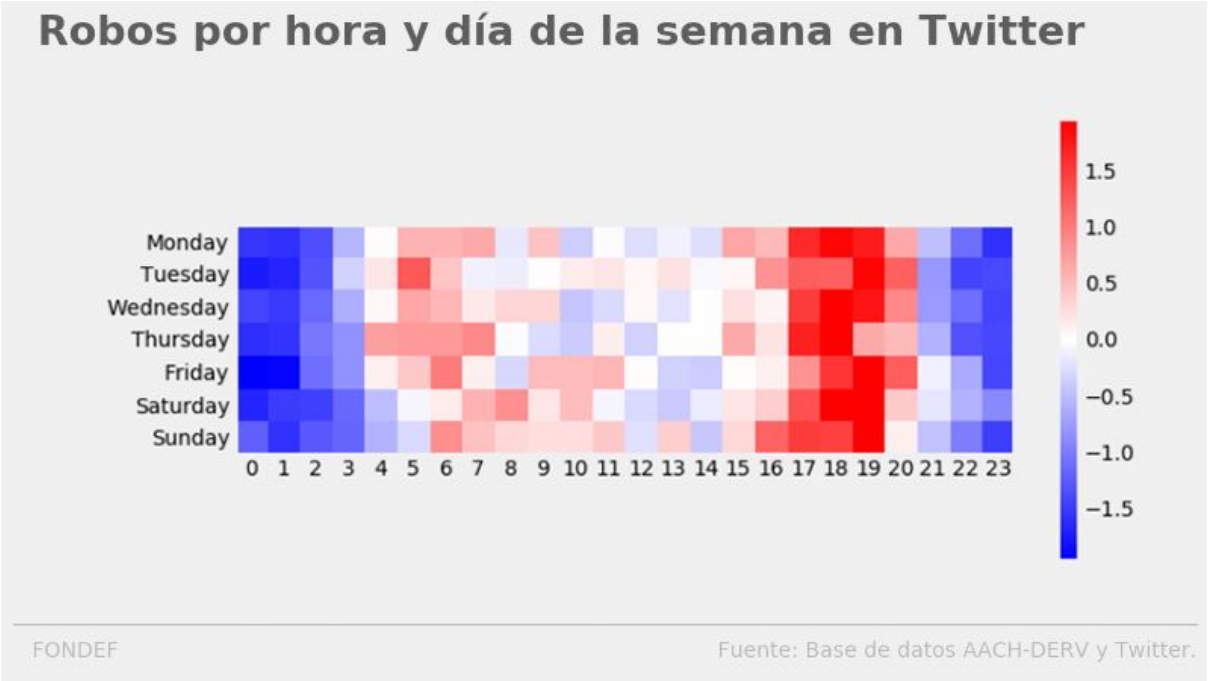

Figura 19. Robos por hora del día y día de la semana en Twitter

En AACH-DERV la mayoría de los casos se concentran en tarde en la noche y temprano en la mañana, mientras que en Twitter se concentran en la tarde, este comportamiento es coherente con que en AACH-DERV se reporta la hora estimada del robo mientras que Twitter es la hora en que se escribe el tweet. Probablemente las horas de la tarde es que las víctimas del robo tienen tiempo para reportar el robo en Twitter. Llama la atención en el caso de AACH-DERV que los fines de semana hay un aumento en la madrugada, mientras que los días de semana este aumento se traslada a las horas de la mañana.

Figura 20. Intervalos de tiempo que transcurren entre las etapas del proceso que comprende el robo de un vehículo y su eventual hallazgo

Para el gráfico anterior se calcula la mediana de los intervalos de tiempo transcurridos entre cada paso por los que transita un robo y el momento del robo. Los casos que se reportan en Twitter ocurren un poco antes de las 20 horas luego del siniestro, entre las 22 y las 25 horas se produce el reporte en carabineros y alrededor de las 95 horas se produce el hallazgo para los casos en que este ocurre.

El siguiente gráfico visualiza las tasas de recuperación para ambas fuentes de datos. Solo tenemos información sobre recuperación para los datos de AACH-DERV, para considerar el impacto de reportar un robo en Twitter comparamos los datos de AACH-DERV cuando se reportan en Twitter con respecto a cuando no son reportados. De los 39.745 casos de AACH-DERV 1.763 son reportados también en Twitter y son recuperados 23.721 y 1.230 respectivamente para unas tasas de recuperación de 0,60 y 0,70, lo que muestra un aumento en el 10% de recuperación para los robos que también son reportados en Twitter.

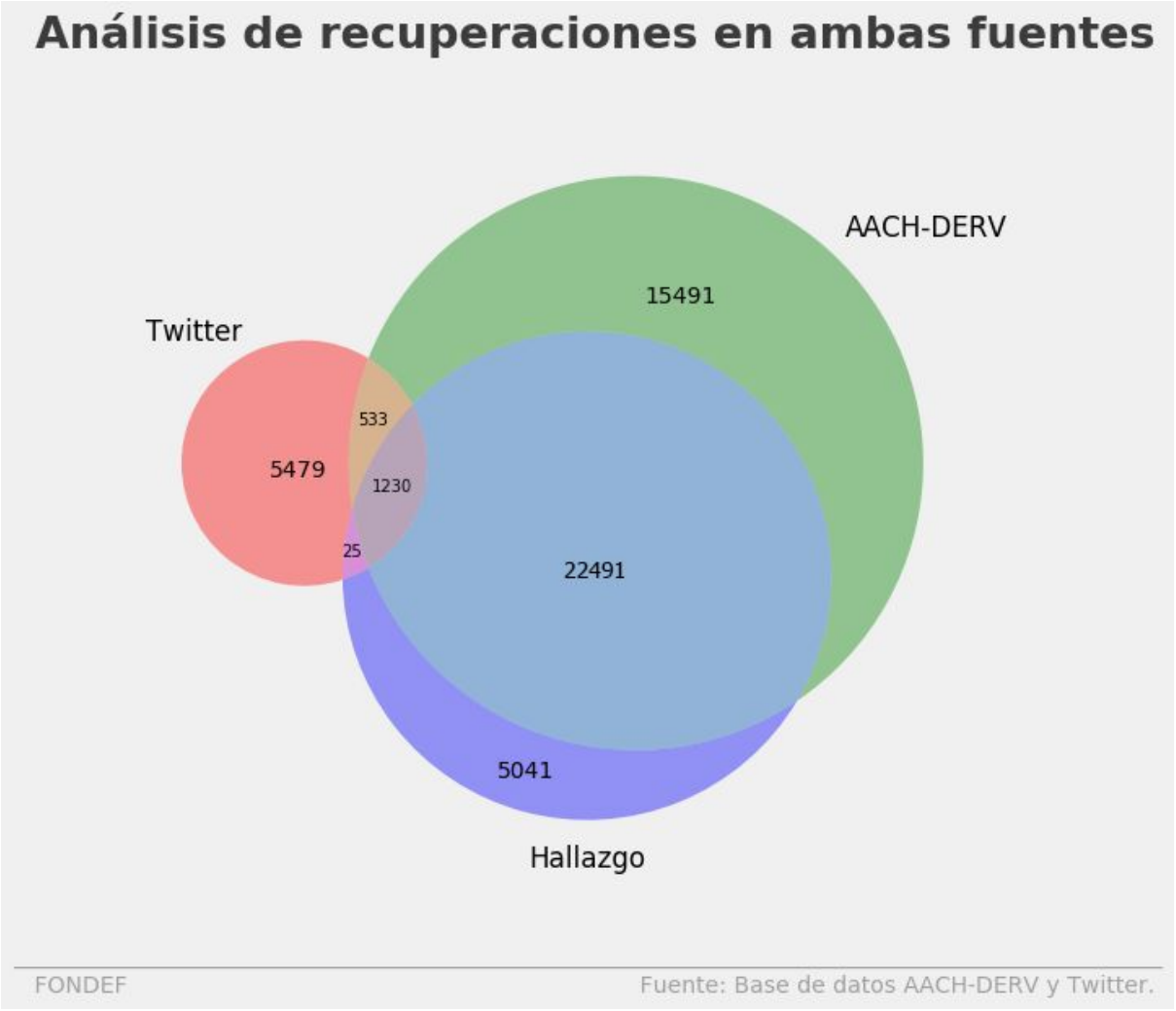

Figura 21. Registros de vehículos recuperados en ambas fuentes de datos

Resultados de consultas sobre accesorios

Además del robo de autos se realizó un estudio exploratorio sobre el reporte de robos de accesorios. Para ello se consultó una base de datos con los tweets pertenecientes al bounding box[2] de Chile recolectado durante el año 2016. Sobre estos datos se consultó:

```
(robar | robado | robaron | robo | robada) & ( llanta | espejos | focos )
```

A continuación mostramos algunos tweets recuperados por esta consulta:

"#BarrioItalia esta muy peligroso ya no se puede ir en auto.Andan robando mal. Me robaron los espejos y aun amigo le rompieron los vidrios"

"Anoche me estacioné afuera del hospital de Iquique me robaron espejo Izquierdo Mazda premacy está con patente GSJD77 si alguien lo ve plis ?"

"@sitiodelsuceso jeep robado hoy en Valparaíso. Es vistoso, tiene focos y una antena con bandera. Cualquier dato sir https://t.co/kXsC4FIFjf"

[2] Concepto geométrico con el que se busca describir un determinado conjunto en n-dimensiones. En este caso representa el conjunto mínimo que describe a Chile en coordenadas de latitud y longitud.

“Me robaron los espejos del auto, pero ya se los repuse”
“3 días se demora el taller de @BciSeguros para hacer el presupuesto de reposición e instalación de los espejos retrovisores que me robaron”
“oigan hoy hay luna nueva, y en las noches de luna nueva no hay que mirarse al espejo o te pueden robar el alma”

Se incluyeron inicialmente estos accesorios porque son los más comunes que se roben en un auto, esta exploración inicial tiene como objetivo descubrir nuevos accesorios, cuando se roban accesorios es probable que se robe otros accesorios que se roban comúnmente. En los tweets recuperados no aparecen otros accesorios reportados salvo *“antena”* . Entre los tweets recuperados el primer conjunto está relacionado con robos de accesorios mientras que el segundo aunque contiene las palabras de la consulta no son reportes de robos de accesorios.

Extracción automática de vocabulario
Con el objetivo de extraer un vocabulario de manera automática para recuperar datos de Twitter asociados a robos de auto que sea capaz de descubrir patrones que evolucionen en el tiempo se realizaron experimentos sobre la extracción automática de vocabulario. La idea es comenzar con un vocabulario inicial que se actualice de manera automática para descubrir patrones nuevos y mejorar la recolección de datos. basándonos en técnicas de representación de textos utilizadas en el área de la minería de textos como TF-IDF y word2vec (Mikolov et al. 2016) . En ambos casos no consideramos stopwords ni palabras con clara vinculación al robo de autos que no nos interesa descubrir como: marcas de autos, colores o conjugaciones de verbo *robar* .

Para aplicar TF-IDF se consideran los tweets asociados a robos recolectados inicialmente como un único documento, como colección de background consideramos tweets extraídos del boundingbox de Chile. El siguiente gráfico muestra los términos con mayor TF-IDF en la colección inicial.

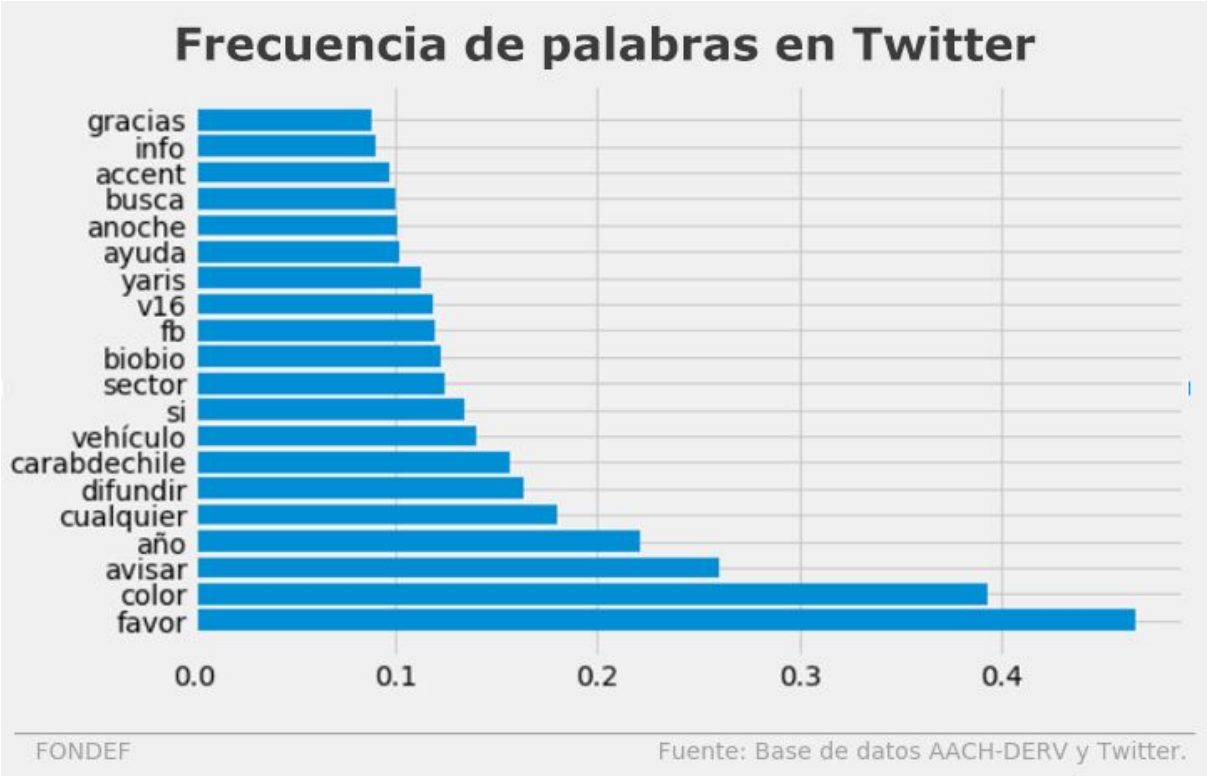

Figura 22. Frecuencia de las palabras que aparecen en los tweets consultados

Word2vec (Mikolov et al. 2016) fue propuesto para obtener una representación vectorial densa para palabras. Los vectores obtenidos presentan propiedades sintácticas y semánticas. En nuestro caso partimos de un modelo de word2vec entrenado sobre la wikipedia en español, luego calculamos vectores para las palabras que aparecen en la colección de tweets asociados a robos promediando los vectores de las palabras que aparecen en un mismo tweet, las palabras que presentan mayor variación con respecto a los vectores del modelo inicial de wikipedia son seleccionadas como las más representativas. El siguiente gráfico muestra los términos con mayor variación con respecto al modelo inicial entrenado sobre wikipedia.

Figura 23. Palabras con mayor variación con respecto al modelo entrenado de forma previa

Conclusiones respecto al uso de Twitter

De manera general se encontraron similitudes y diferencias entre ambas fuentes de datos. Entre las similitudes podemos destacar los patrones temporales, existe bastante correlación en cuanto al volumen de robos reportados en AACH-DERV y Twitter y en los años recientes es aún mayor. La proporción de robos por dia de la semana es similar en ambas fuentes. Entre las diferencias podemos destacar que la tasación de los autos reportados en Twitter es inferior a la de los autos en AACH-DERV. En cuanto a las marcas de auto las proporciones son similares, pero si consideramos los modelos vemos que algunos modelos tienden a aparecer en mayor frecuencia en una fuente de datos antes que en otra, por ejemplo el Nissan V16, el cual es utilizado preferentemente para el transporte de pasajeros, los cuales no son asegurados, y por ende la presencia en los datos de AACH-DERV es menor.

Otro aporte de este trabajo es que se recreó el proceso por los cuales pasa un robo de vehículo, en donde se descubrió que el orden es Robo del vehículo, Envío del Tweet, Denuncia en Carabineros de Chile, Hallazgo. El tweet es el primer evento que se origina luego del robo.

El tweet se genera antes que la denuncia en Carabineros de Chile, y con varios días de anticipación de la obtención de los datos por parte de AACH-DERV. Además se evidenció que los vehículos denunciados por robo en Twitter presentan mejor tasa de hallazgo que aquellos en que no se realiza la denuncia por esta red social.

5.1.3. Noticias

El objetivo de esta sección es la validación de las noticias digitales como una fuente de información apropiada para el desarrollo del observatorio. Para este fin se realiza en primera instancia el levantamiento de una base de datos con noticias publicadas en importantes medios digitales, para luego realizar un análisis prospectivo de estas noticias. Basados en este análisis se busca validar o descartar la significancia de este medio como fuente de información confiable con respecto al robo de automóviles en Chile.

Una vez obtenida la base de datos (cuyo proceso de elaboración se aborda en los siguientes apartados) se estudia el cuerpo de noticias mediante técnicas de text mining. Los campos considerados por noticia son: fuente (de la noticia), titular (de la noticia), texto (cuerpo de la noticia), fecha de publicación (de la noticia), imágen (de la noticia).

5.1.3.1. Construcción y descripción general de la base de datos de noticias digitales

Para la construcción de la base de datos se privilegió el uso de información pública, de modo de asegurar el funcionamiento del observatorio con información sin ningún tipo de restricción de acceso. Dado lo anterior, la base de datos fue construida con noticias publicadas y dispuestas a disposición pública por los medios de prensa digitales abiertamente en sus sitios web.

La captura de la información se realiza mediante la obtención del código fuente la página web que contiene la noticia (incluyendo url de imágenes, multimedia y material escrito). Lo anterior se hace mediante la consulta del código fuente de la página web (HTML) y luego analizando este código para extraer la información relevante (a este proceso se le llama "*scraping*"). El mayor desafío en el "*scraping*" de noticias es la heterogeneidad de las páginas web que contiene la información. La estructura de cada página (y del código fuente) varía con cada proveedor. Dado esto, la captura de noticias se realizó usando software construidos de manera ad-hoc para cada sitio de noticias en particular.

A pesar de lo heterogéneo de cada fuente específica de información, la metodología de captura de datos es la misma para cada sitio y consistió de tres pasos:

1) Scrapeo de titulares y urls de noticias desde la página principal del sitio web.
2) Scrapeo de la información específica de noticias desde la url capturada en el paso 1.
3) Almacenamiento de la información en una base de dato relacional con los siguientes atributos por noticia: medio escrito, titular, cuerpo, fecha de publicación, region de publicación, imagen principal.

Independientemente del medio, todos presentan un listado de noticias en la forma de una página principal (http://lanacion.cl/ por ejemplo). Desde ésta página principal (la llamaremos "portal principal" de aquí en adelante) se puede extraer un número de links específicos (http://lanacion.cl/2017/09/25/los-principes-de-japon-llegan-este-martes-a-santiago/ por ejemplo). En la primera etapa de nuestra metodología estos links son almacenados temporalmente para luego analizarlos específicamente y extraer la información de interés

para nuestro estudio. Concretamente para bajar el listado de links a noticias desde la página principal se utilizó el browser lynx y una serie de filtros de texto para extraer los links relevantes a nuestro análisis (el comando grep específicamente). Este proceso de extracción de noticias para la página principal se puede extender a otras pagina en el historial del sitio (http://lanacion.cl/page/2/ para ir a otro portal con mas links) o analizando otros links dentro del portal.

Una vez construido el listado se procede a analizar el código HTML de cada página para rescatar la información relevante en cuanto a urls y texto. Para esto se utiliza un script que obtiene el código fuente HTML de la página mediante el programa wget y luego lo analiza con un script escrito en el lenguaje computacional python (con el paquete BeautifulSoup). Mediante este software se puede filtrar información específica (por ejemplo el marcador de título en HTML “title”), la que es luego rescatada, estructurada de acuerdo a los campos definidos de la base de datos y guardada localmente. Con este último archivo almacenado se pasa a el almacenamiento de esta información en la base de datos. En nuestro caso el almacenamiento se realizó en una base de datos relacional SQL, particularmente se usó el softwares PHPMyadmin para la administración de la base de datos SQL.

La ejecución automatizada de los pasos de la metodología descrita se realizó mediante un script en lenguaje shell script (.sh). Este script ejecutaba secuencial y ordenadamente los pasos de obtención y almacenamiento de links desde el portal principal, análisis individual y consulta del código fuente de cada link almacenado, filtrado, extracción y almacenamiento temporal de la información necesaria y finalmente limpieza de los repositorios temporales locales.

Mediante la aplicación de la metodología descrita se recabaron 9.430 noticias publicadas en los sitios web bio-bio.cl, la Nación.cl y Emol.cl entre Junio de 2016 y Junio de 2017. La siguiente figura presenta la distribución de la cantidad de las noticias consideradas en función de su fecha de publicación. La distribución en la cantidad de noticias extraídas desde bio-bio.cl y emol.cl es homogénea para todos los meses, a diferencia de la Nacion.cl, que se concentra en los meses de Mayo y Junio de 2017. Por este motivo se realiza un análisis diferenciado de este medio, sin encontrar diferencias relevantes con respecto a los resultados generales.

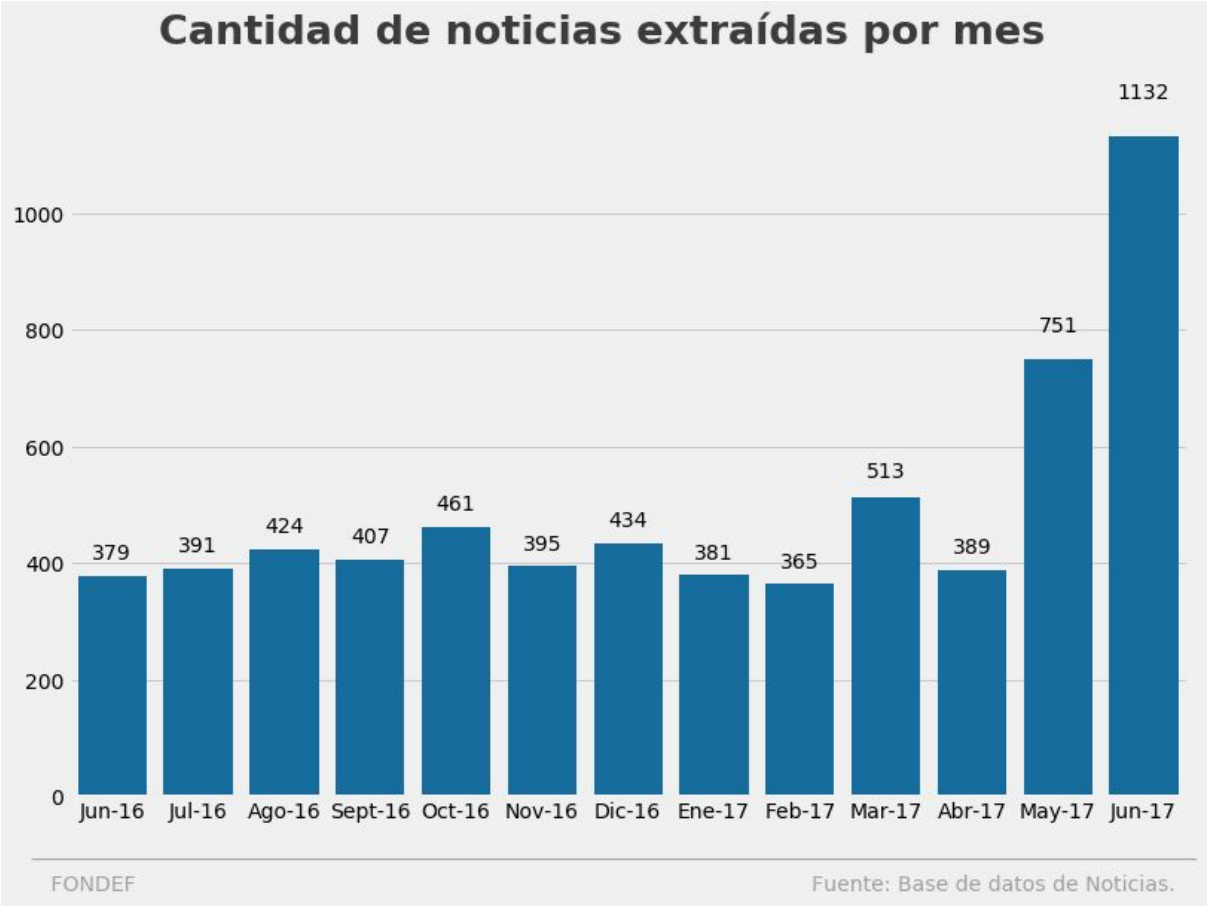

Figura 24. Distribución de noticias por mes

5.1.3.2. Análisis de tópicos

Para explicitar el contenido de las noticias, se realiza un análisis de tópicos. Este análisis es un proceso iterativo de clasificación no supervisada de noticias. En particular, se utiliza la técnica Latent Dirichlet allocation (LDA) en donde cada noticia es un documento y cada documento contiene una mezcla de tópicos.

De este análisis se obtienen 6 tópicos junto a la probabilidad de cada documento con respecto a los tópicos. Los tópicos son los siguientes (de acuerdo al tipo de noticia predominante): 1) Autoridades – Política, 2) Educación, 3) Policial – Delitos, 4) Economía, 5) Accidentes vehiculares – Atropellos y 6) Nacional.

El tópico 6, Nacional, es un conjunto de varios temas de interés noticioso entre los que destacan: Meteorología y Condiciones medioambientales; Cortes del suministro de luz y agua; Relacionadas con salud; Emergencias (sismos, incendios, lluvias); Transantiago y Metro; Relacionadas con problemas laborales.

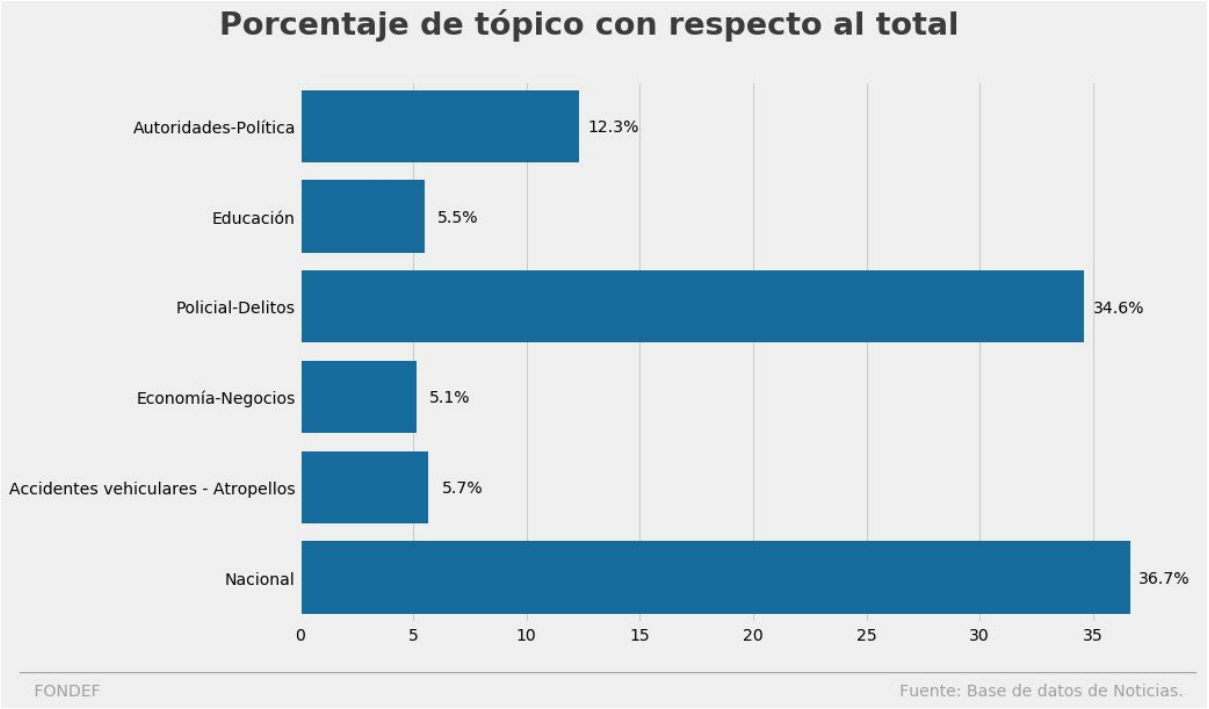

Figura 25. Porcentaje de noticias por cada tópico

Las principales palabras por cada tópico son las siguientes:

Tópicos	Palabras relacionadas
Autoridades-Política	Alcalde – Alessandri – Municipio – Autoridades – Ministros – Bachelet – (Pre) Candidato – Candidatura – Primarias – Lavín – Concejales – Intendencia – Comisión – Diputado – Senador – Piñera – Seremi
Educación	Liceo – Apoderado – Estudiantes – Universitarios – Colegio – Universidad – Marcha – Secundario – Estudiantil – Confech - Rector
Policial-Delitos	Acoso – Allanamiento – Querella – Demanda – Maltrato – Sumario – Abuso – Apuñalan – Asaltante – Asalto – Delincuente – Balacera – Baleado – Disparo – Banda – Ataque – Carabineros – Tribunal – Encapuchados – Víctima – Condenan – Culpable – Prisión – Presidio - PDI
Economía-Negocios	Inversión – Déficit – Deuda – Crédito – Empresa – Concesión – Quiebra – Financiero – Pesos – Millones – Cobro – Morosos - Evasión
Accidentes vehiculares	Vehículo – Automóvil – Atropello – Colisión –

– Atropellos	Choque – Bus – Camión
Nacional	Agua, Luz, Meteorología, Medioambiente, Bombero, Incendio, Sismo, Transantiago, Metro, Recorrido, Huelga, Trabajador

Tabla 2. Principales palabras por tópico identificado

Finalmente, el objetivo de este proceso es aislar las noticias que corresponden a delincuencia, para luego, detectar las noticias correspondientes a robo de automóviles.

5.1.3.3. Tópico Policial – Delitos

Más de un tercio de las noticias están asociadas a este tópico, siendo las noticias de delitos las predominantes. Keywords como Delincuente, Asalto, Robo Crimen, Detenidos, Condena, Prisión tienen la mayor probabilidad de aparecer.

Al realizar un análisis de tópicos solo de este tipo de noticias, se tiene que hay 6 subtópicos predominantes: 1) Asalto, 2) Agresión – Crimen, 3) Detención – Sentencias, 4) Robo de automóviles, 5) Acoso y 6) Otros.

En la tabla 3 se muestra el porcentaje de noticias correspondiente a cada subtópico dentro de Policial – Delitos.

Tópico	% respecto al total Policial-Delito
Asalto	26.03%
Agresión-Crimen	23.52%
Detención-Sentencias	20.58%
Robo de automóviles	8.69%
Acoso	5.88%
Otros	~15.3%

Tabla 3. Porcentaje de noticias de cada subtópico

Dentro de este tópico el 8.69% de las noticias están relacionadas al robo de automóviles (2.09% del total de noticias, considerando todos los tópicos), siendo el cuarto subtópico en orden de importancia.

En particular, para el subtópico Robo de automóviles, interesa conocer el tipo de noticias que se publican. Se tiene entonces que:

- El 35% de estas noticias corresponden a Detenciones o Sentencias, a Asaltantes o Bandas Criminales de robo de vehículos
- El 20.33% corresponde a la recuperación de vehículos robados
- El 30.88% reporta el robo de automóviles

En el último punto se tienen las noticias directamente aplicables al proyecto, esto es, 1.16% del total de noticias.

Las noticias que se publican son, en su mayoría, relacionadas con la "espectacularidad" del hecho (ej. Persecuciones, autos de lujo, etc.) o debido a la relevancia de la víctima (famoso de TV, deportista, político, etc.). Ejemplos:

- "Al menos 30 vehículos robados fueron recuperados en sitio eriazo de Maipú". 9-05-2017.
- "Apuñalan a rector de la U. Iberoamericana tras robo frustrado de su auto en Las Condes". 10-05-2017
- "Carabineros recupera lujoso vehículo robado a mujer en Vitacura". 3-06-2017
- "Detienen a 3 adolescentes que integraban banda especializada en portonazos de Santiago". 26-04-2017
- "Marcelo Salas sufrió el robo de su Porsche en Vitacura". 18-04-2017

5.1.3.4. Relevancia del medio y conclusiones

Del total de noticias correspondientes al subtópico Robo de Automóviles, sólo en el 25% de las noticias se entrega información de la marca del vehículo, y detalles del robo como la fecha y el lugar. Esto es, en el 0.75% de las noticias totales se cuenta con detalles relevantes para este proyecto. Se concluye entonces que la cantidad de noticias publicadas con respecto al robo de automóviles no es suficiente para establecer ningún tipo de relación causal o para denotar a este medio como un termómetro de la realidad país en el tema.
Más aún, al comparar los datos de noticias con la base de datos de AACH-DERV, se destaca la diferencia entre ambas fuentes de información. No es posible observar tendencias similares ni obtener correlaciones con el fin de respaldar o complementar información.

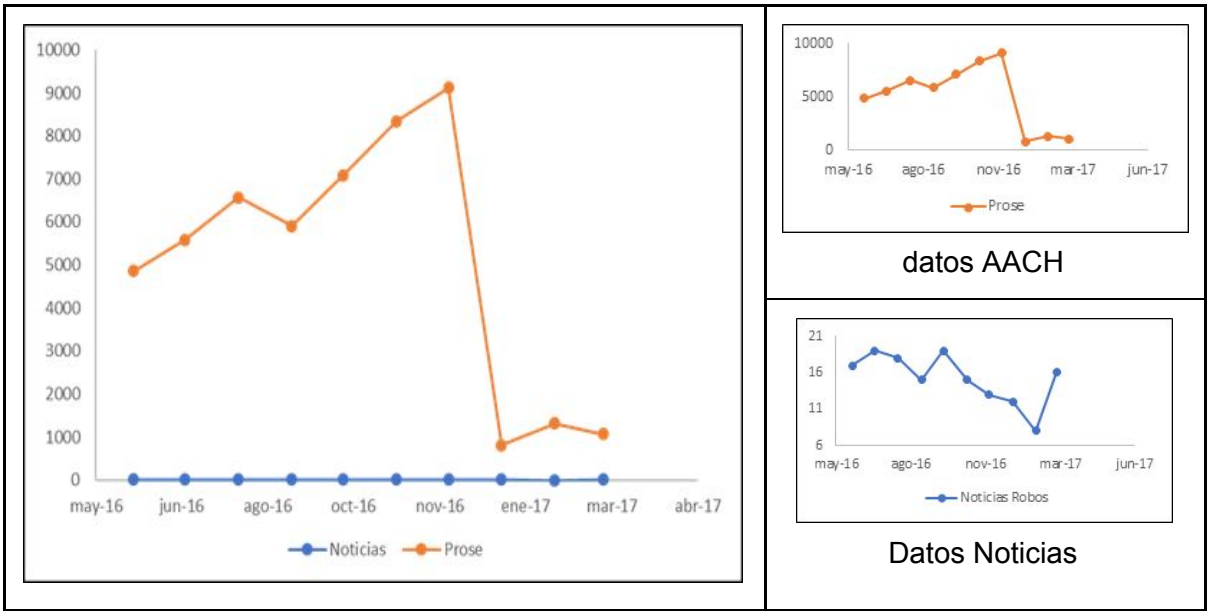

Figura 26. Serie de tiempo comparando volumen de datos de AACH con datos de noticias

Mes	jun-16	jul-16	ago-16	sep-16	oct-16	nov-16	dic-16	ene-17	feb-17	mar-17
Noticias	17	19	18	15	19	15	13	12	8	16
AACH-DERV	4860	5586	6576	5904	7090	8338	9120	819	1318	1071

Tabla 4. Volumen de registro de robos de vehículos en ambas fuentes de información

5.2. Análisis exploratorio

5.2.1. Análisis por tipo, marca y modelo

En la siguiente se muestra la proporción de los robos registrados según el tipo de vehículo al cual corresponde el automóvil sustraído, siendo estos los 8 tipos de vehículos con mayor presencia en la base de datos. Cómo se puede apreciar, Station Wagon y Automóvil (por automóvil entiéndase un sedan, hatchback, cross over, coupe, etc.) registran cerca del 73% de los casos, siendo el tercer vehículo con mayor registro las Camionetas. Estos vehículos corresponden en su mayoría a medios de transporte privado no comercial.

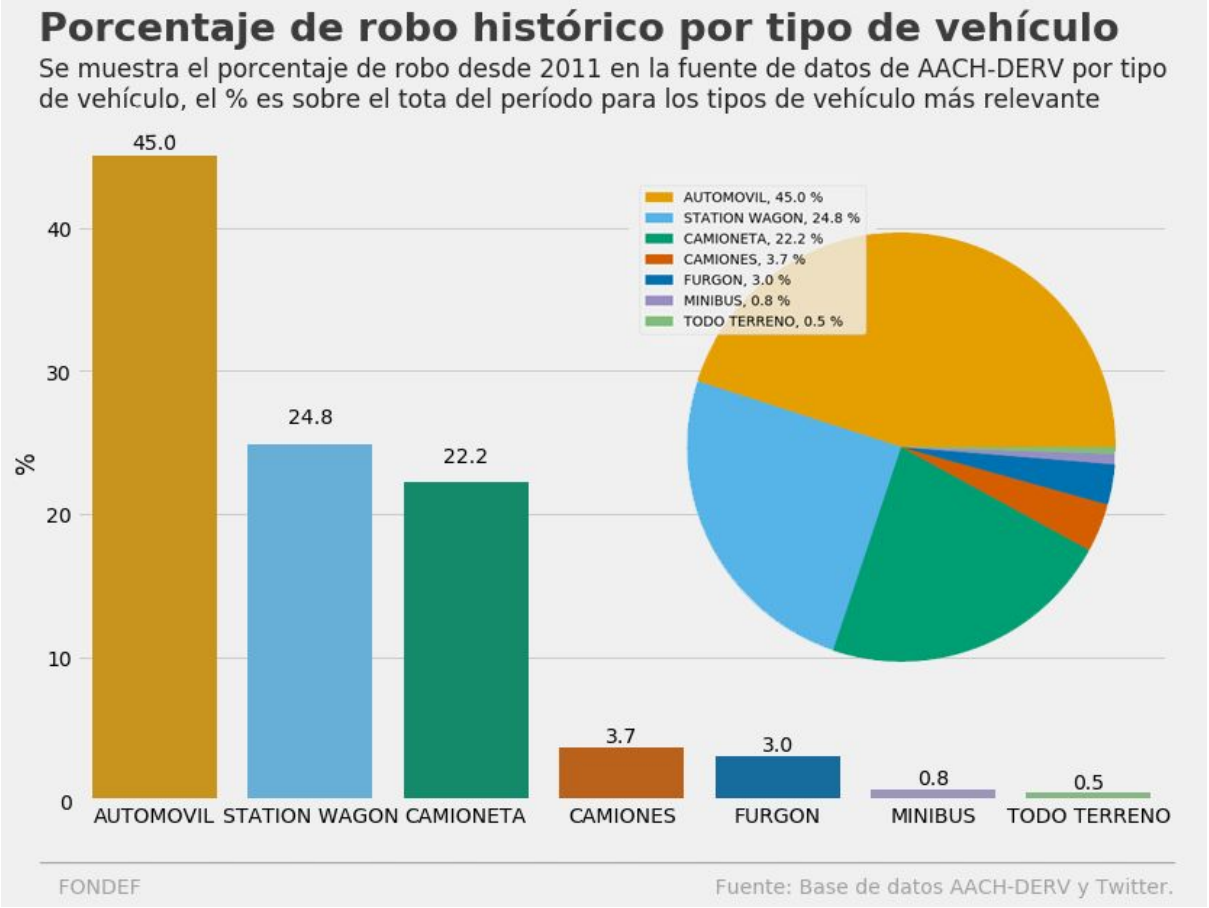

Figura 27. Proporción de los robos por tipo de vehículo

La Figura 28 muestra la composición del robo de vehículos según su tipo distinguiendo entre la Región Metropolitana y las otras 14 regiones. Se puede observar que las diferencias entre la Región Metropolitana y el resto no son sustanciales, lo cual permite

concluir que el robo de este tipo de vehículos está determinado por su proporción en el mercado más que por algún tipo de atractivo en su robo, es decir, corresponde a un efecto sistemático más que a un efecto de preferencias por los delincuentes.

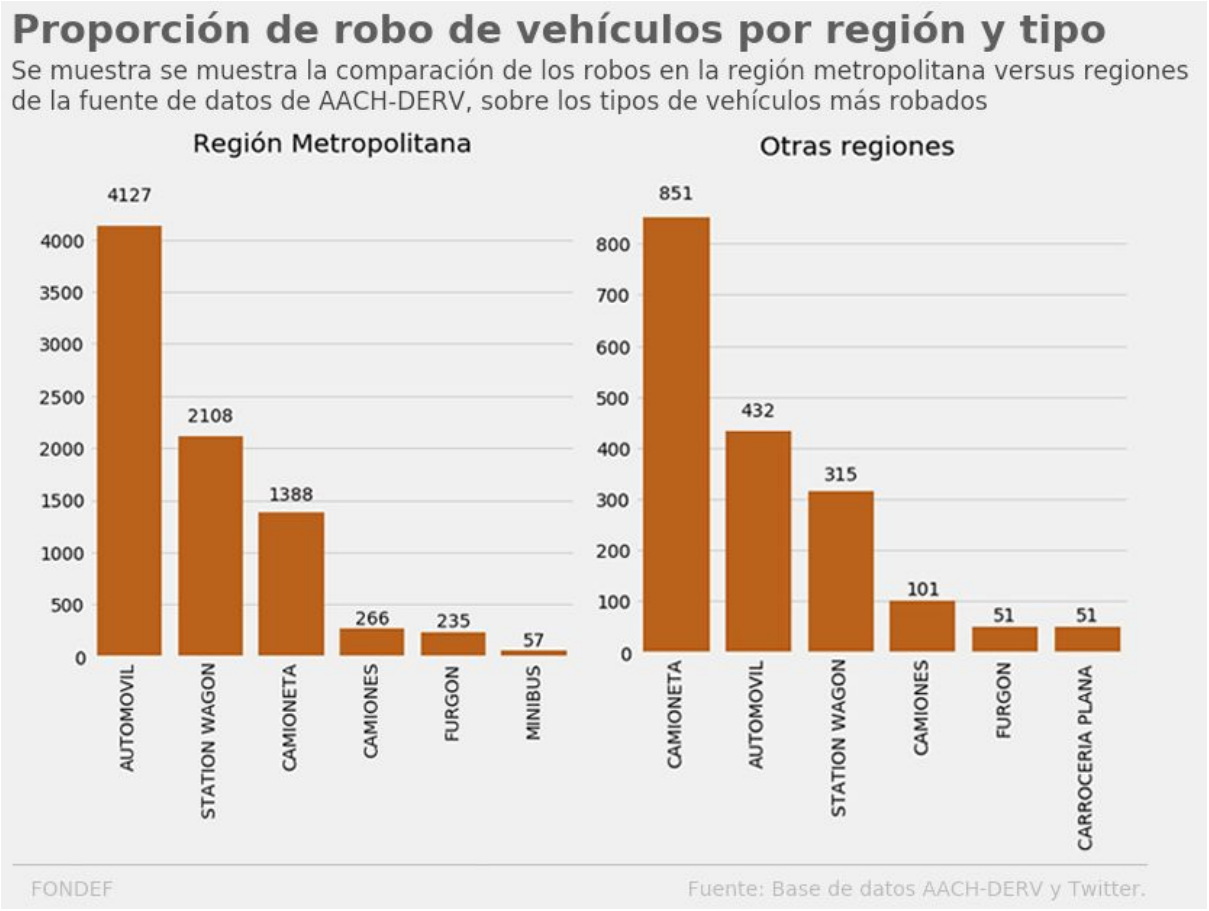

Figura 28. Descomposición por regiones de la proporción de los robos por tipo de vehículo

En la Figura 29 se puede analizar la descomposición temporal del robo de vehículos según su tipo. Como se puede apreciar, estos han seguido una tendencia bastante regular, manteniéndose constante a través del tiempo las proporciones del tipo de vehículos más robados. Nuevamente se puede hipotetizar que el robo de un determinado tipo de vehículo está primeramente determinado por su presencia en magnitud en el mercado.

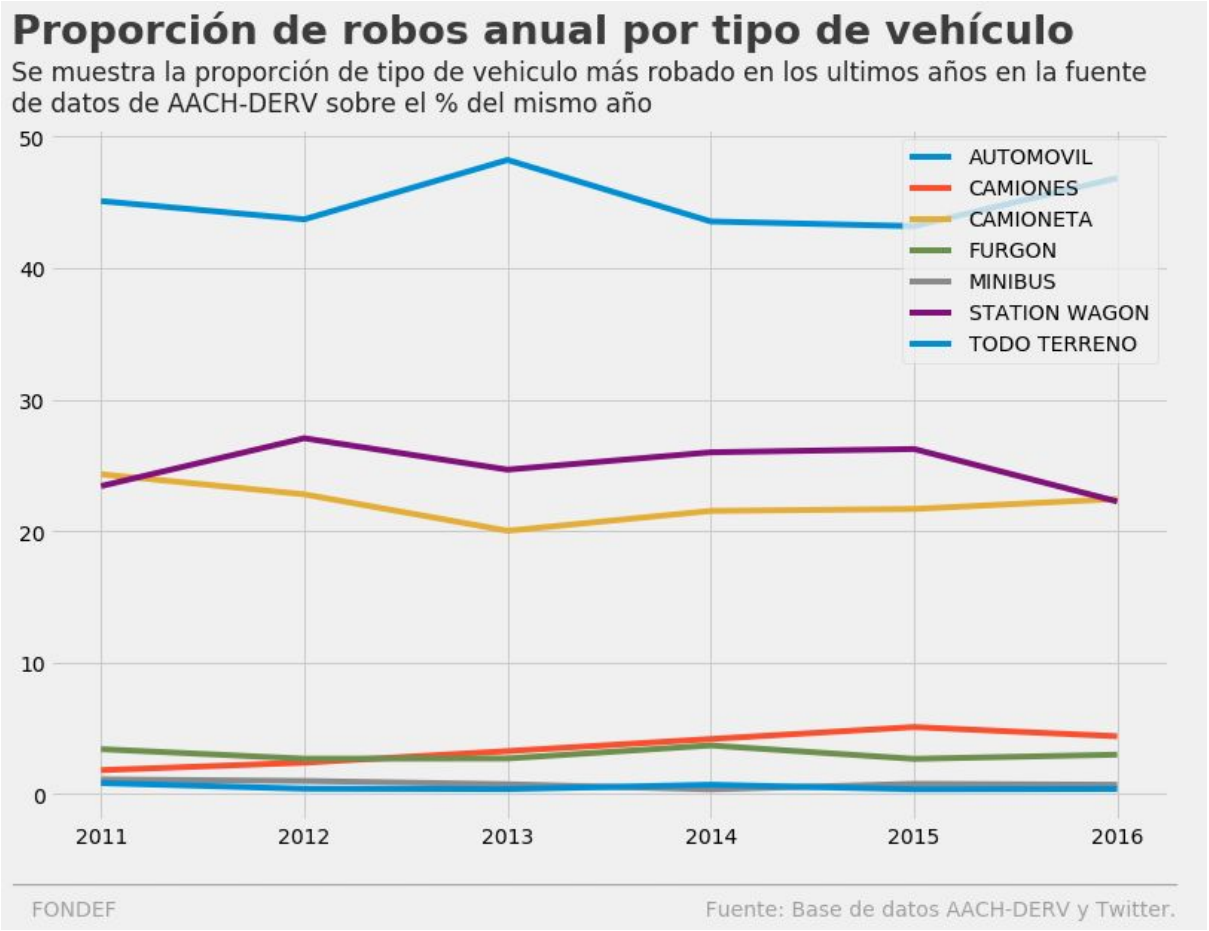

Figura 29. Análisis temporal de la proporción del robo de tipo de vehículos

En la Figura 30 se ha realizado un análisis que muestran la proporción del robo de vehículos según su tipo, el lugar donde fueron sustraídos y todo estos diferenciando por Región Metropolitana o regiones. Los resultados son consistentes con los resultados anteriores, en cuanto se mantiene la proporción del robo de vehículos según su tipo, lo que quiere decir que sistemáticamente el Automóvil, las Camionetas y Station Wagons son los vehículos más robados, a la vez que en la Región Metropolitana la calle o vía pública es el lugar donde más se sustraen vehículos, mientras que en regiones se invierte esta proporción, siendo el domicilio el lugar con mayor proporción de robos. No obstante, sea calle o domicilio, las proporciones por tipo de vehículo se mantienen.

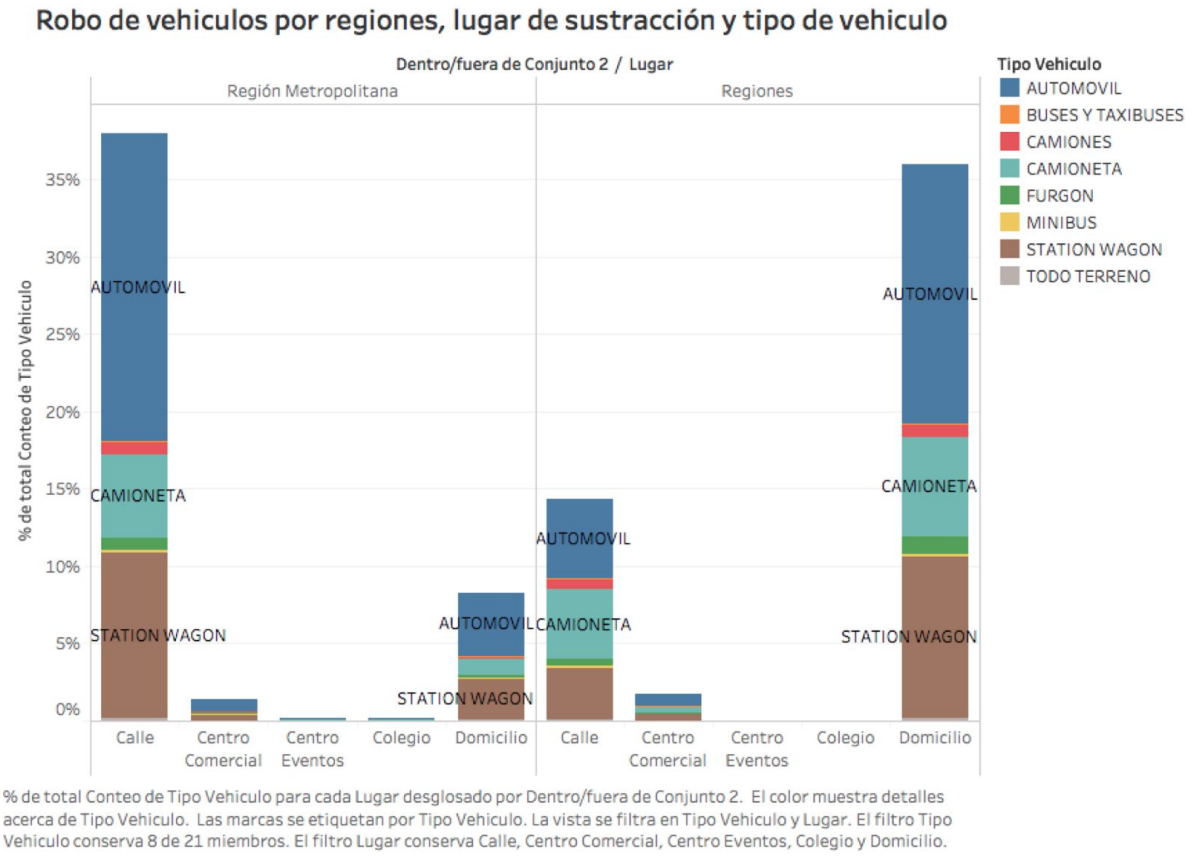

Figura 30. Análisis según el tipo de vehículos y el lugar donde son sustraídos por región

En la Figura 31 se observan las 12 marcas con mayor proporción de robos en Chile (por simplicidad algunas de ellas han sido agrupadas), las cuales sumadas representan más del 85% de los robos registrados en el periodo 2011-2016.

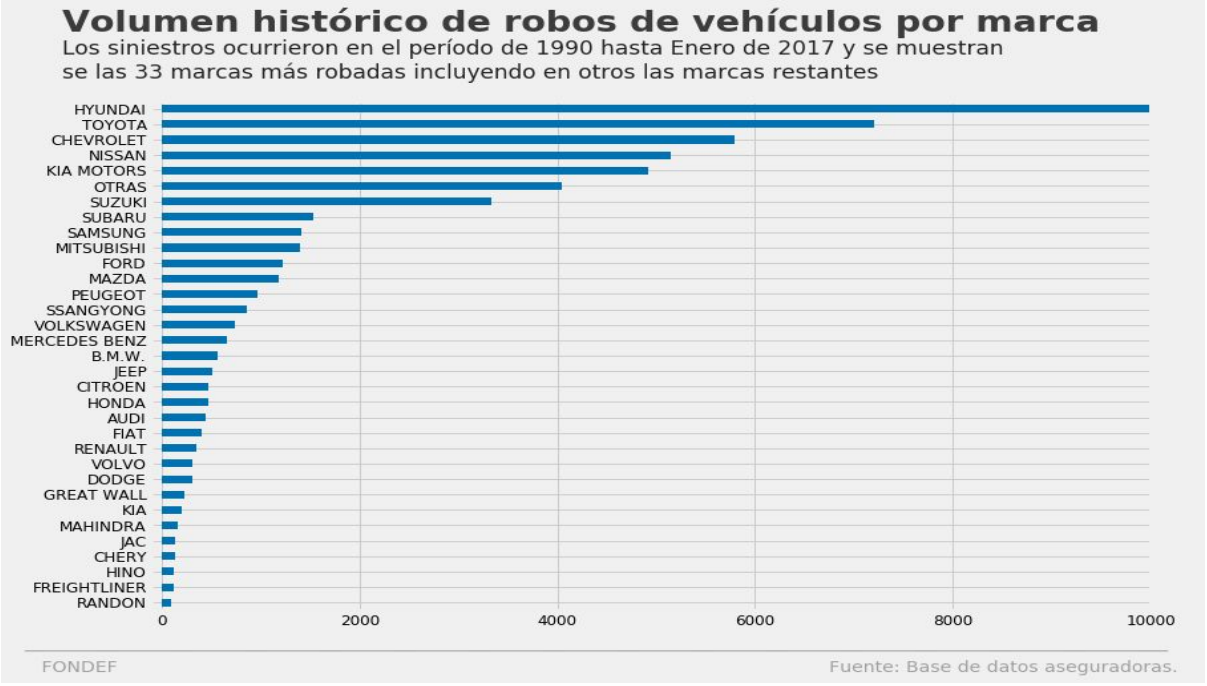

Figura 31. Marcas de vehículos con mayor cantidad de robos

En este punto es necesario reflexionar respecto a porque son estas las marcas de vehículos más robadas ¿Será posible que su bajo nivel de tecnología haga a sus vehículos más vulnerables? En la tabla 5 se contrasta el porcentaje de la participación de mercado reportado por ANAC (Asociación Nacional Automotriz de Chile) en su Anuario Automotriz 2015/2016 respecto al año 2014 con la cifra de robos registrada por AACH en el mismo año.

Marca	Participación de mercado	Porcentaje de robos registrados
Chevrolet	14,0%	8,4%
Hyundai	9,8%	22,9%
Kia	9,4%	7,9%
Toyota	6,9%	17,7%
Suzuki	6,5%	4,6%
Nissan	6,2%	9,8%
Mitsubishi	4,2%	2,6%
Ford	4,1%	2,1%
Peugeot	3,5%	1,3%
Mazda	3,4%	2,2%

Tabla 5. Participación de mercado y porcentaje de robos de las 10 principales marcas

De la tabla anterior se puede apreciar que las marcas Hyundai y Toyota están sobrerrepresentadas en la cantidad de robos registrados vs su participación de mercado. No obstante, se puede apreciar que este fenómeno probablemente se debe a que estas marcas poseen a 3 de los vehículos más robados: Toyota Hilux, Hyundai Santa Fe y Hyundai Tucson, los cuales a su vez son vehículos muy populares.

En la Figura 32 se observa la evolución en el tiempo de la cantidad de robos que se han registrado sobre estas marcas.

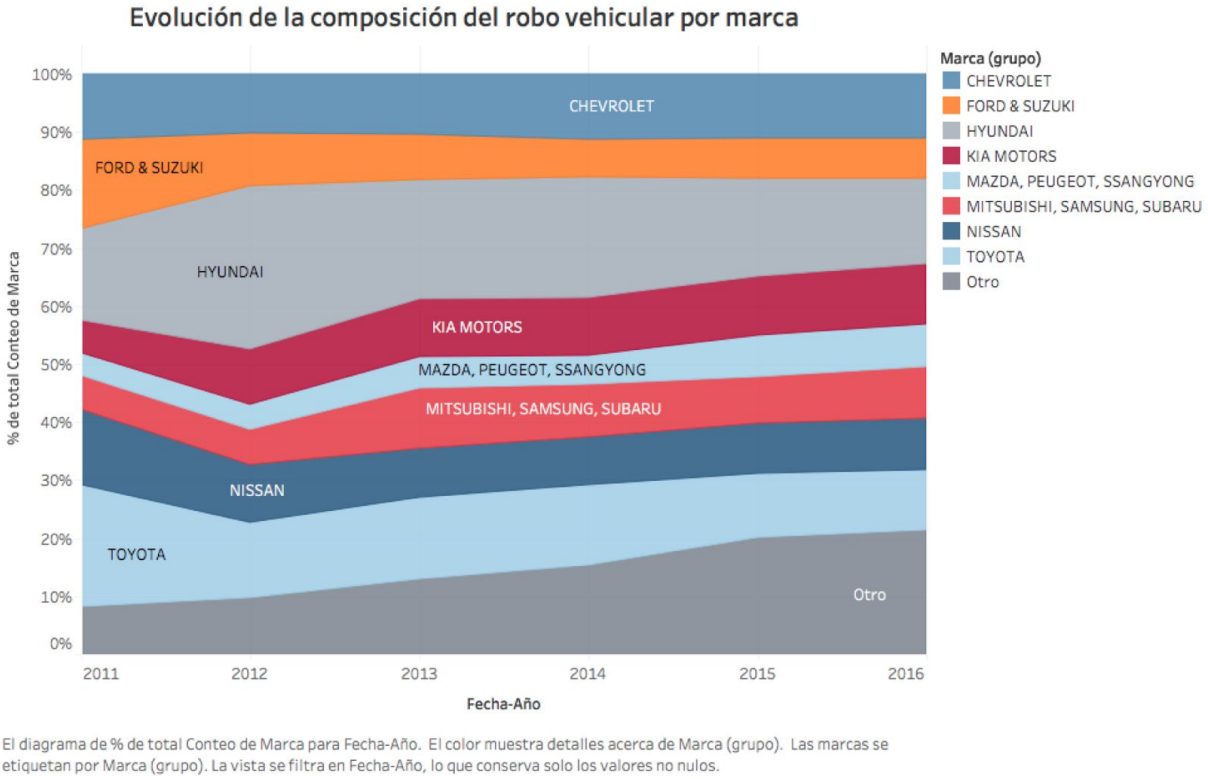

Figura 32. Evolución de las marcas de vehículos con mayor cantidad de robos

La Figura 33 muestra la proporción de los 10 vehículos más robados entre el año 2011 y 2016. Contrario a lo que sucede con las marcas de los vehículos, los 10 vehículos más robados no suman el 20% del universo. No obstante esta cifra sigue siendo significativa, ya que el universo de modelos registrados en la base de datos alcanza los 3248 modelos (aunque también se incluyen modelos de vehículos no destinados a transporte privado, como como maquinaria, vehículos de transporte pesado, vehículos náuticos, etc.)

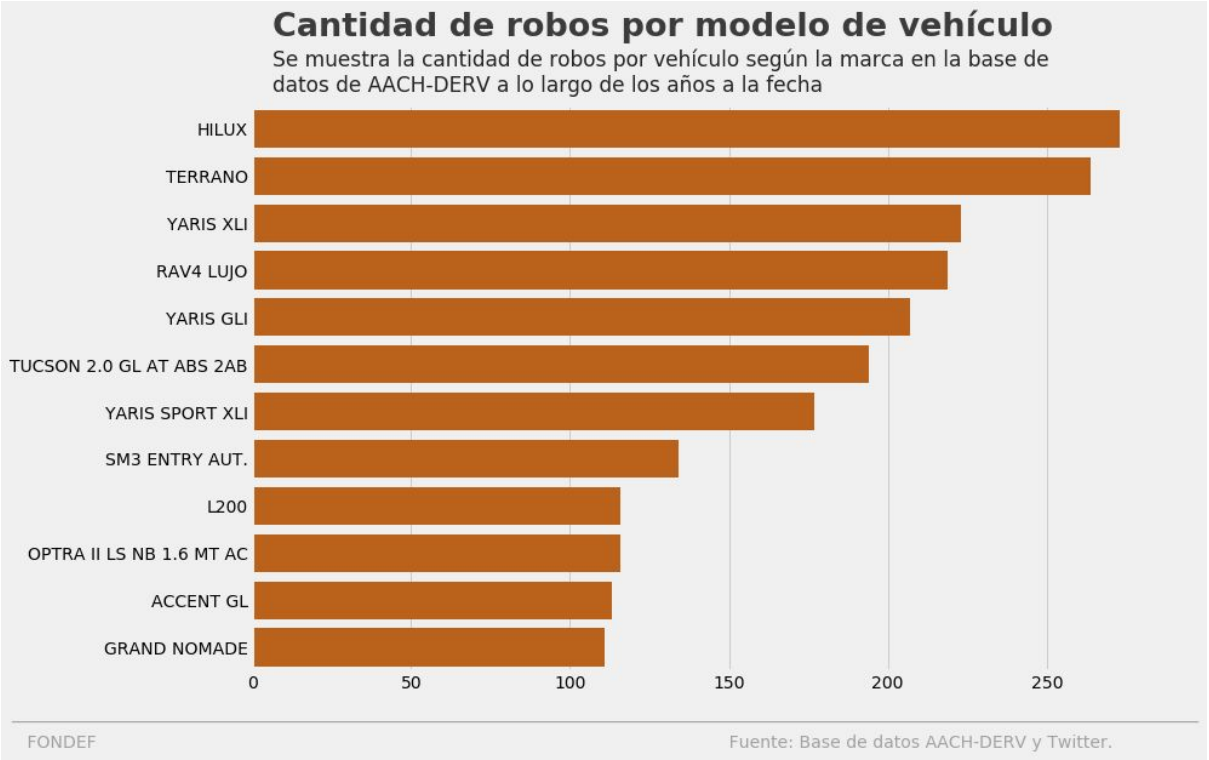

Figura 33. Análisis de la proporción de los vehículos más robados

En las Figuras 34 y 35 se detalla la evolución del robo de vehículos por modelo, especificando las cifras para los 10 modelos de vehículos más robados. Se puede observar como la Toyota Hilux es el vehículo que ha experimentado la mayor alza desde el 2011, alcanzando un 4% de todos los robos registrados en el 2016. Así mismo, se puede observar cómo en comparación al año 2012, en el 2016 los top 10 alcanzaron su mínimo representación en la población de vehículos robados.

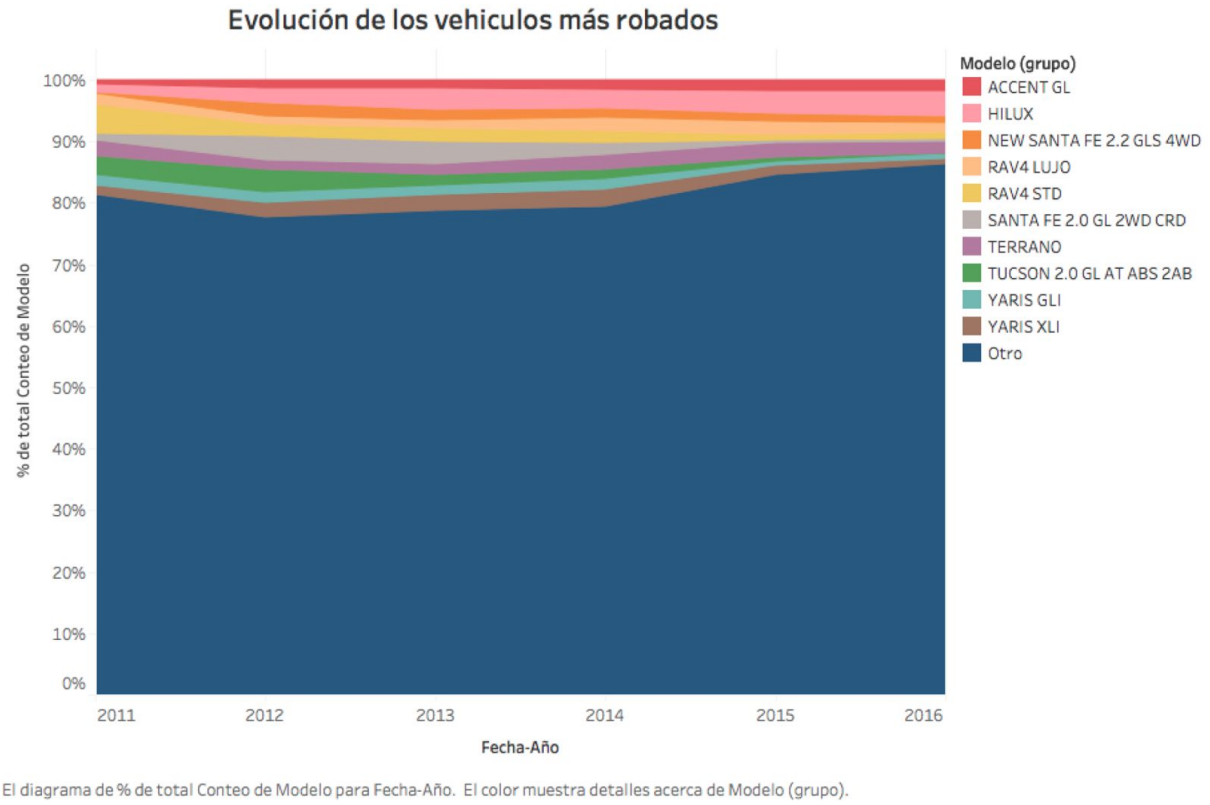

Figura 34 Evolución temporal de los vehículos más robados

Tabla de modelos de vehiculo más robados

Modelo (grupo)	2011	2012	2013	2014	2015	2016
ACCENT GL	0,67%	1,35%	1,25%	1,57%	1,74%	1,74%
HILUX	1,17%	2,37%	3,47%	2,93%	3,70%	4,05%
NEW SANTA FE 2.2 GLS 4..	0,33%	2,13%	1,76%	1,57%	1,25%	1,01%
RAV4 LUJO	1,75%	1,37%	1,28%	2,24%	2,17%	1,53%
RAV4 STD	4,79%	1,76%	2,12%	1,84%	0,93%	1,11%
SANTA FE 2.0 GL 2WD CRD	1,00%	3,90%	3,74%	2,02%	0,40%	0,44%
TERRANO	2,64%	1,67%	1,84%	2,35%	2,44%	1,96%
TUCSON 2.0 GL AT ABS 2AB	3,08%	3,61%	1,55%	1,40%	0,54%	0,06%
YARIS GLI	1,57%	1,76%	1,62%	1,90%	0,65%	0,86%
YARIS XLI	1,55%	2,29%	2,57%	2,65%	1,52%	0,78%
Otro	81,45%	77,79%	78,81%	79,55%	84,67%	86,46%

Fecha-Año

Fecha-Año para cada Modelo (grupo). Las marcas se etiquetan por % de total Conteo de Modelo. La vista se filtra en Fecha-Año, lo que conserva solo los valores no nulos.

Figura 35. Evolución temporal de los vehículos más robados (cifras exactas)

5.2.2. Análisis por tiempo

En la Figura 36 se puede apreciar cómo se distribuye el porcentaje de robos que se registran durante el año en los respectivos meses. El principal hallazgo es que en el mes de Septiembre se registra la tasa más baja de robo de vehículos, fenómeno que no posee explicaciones económicas, salvo el posible aumento del resguardo policial en las calles producto de las fiestas patrias.

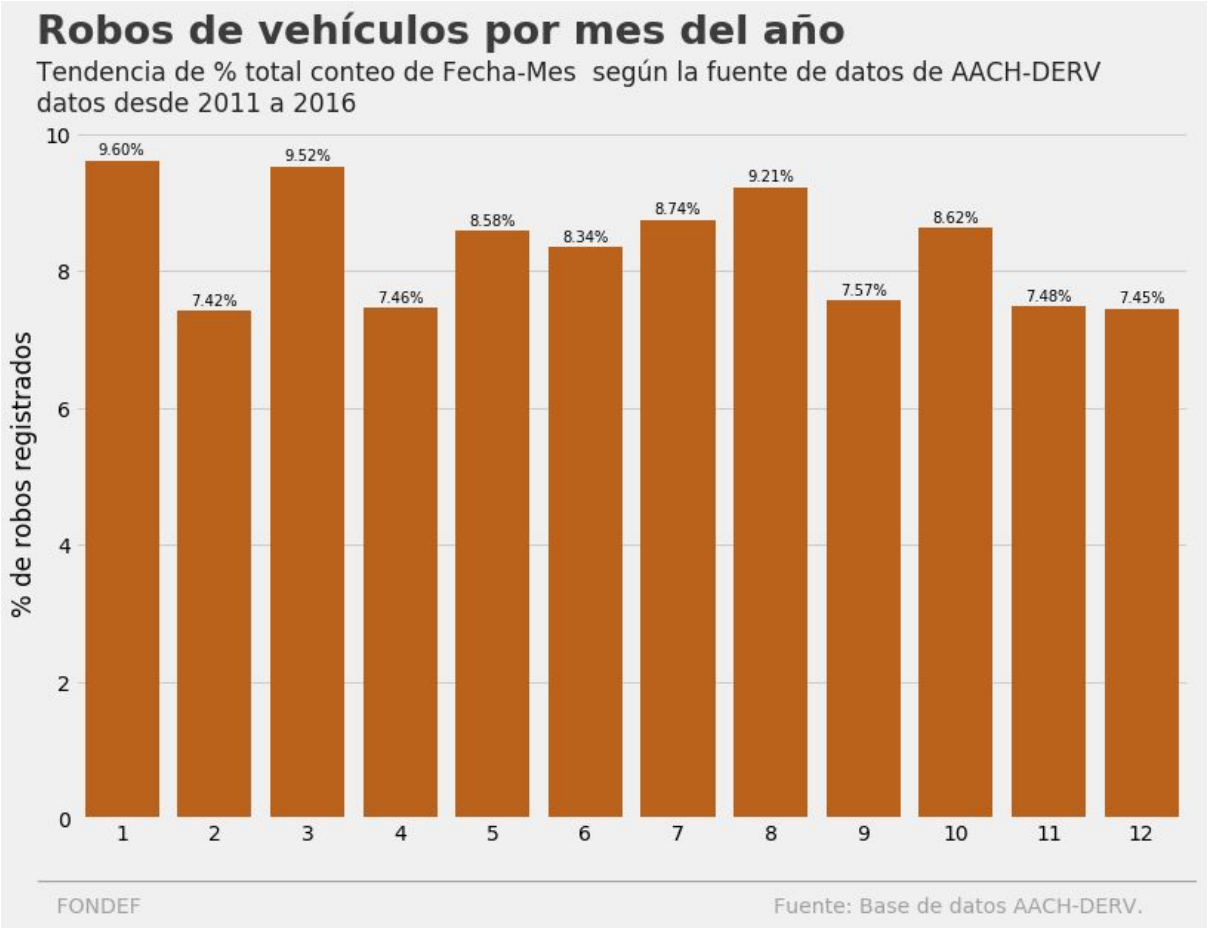

Figura 36. Robo vehicular por meses del año (desde 2011 a 2016)

5.2.3. Análisis por lugar

En la Figura 37 se muestra la proporción global de los lugares donde han sido registrados los robos de los vehículos. Cabe destacar que esta imagen refleja la cifra sin discriminar por el año donde se ha cometido o la región del suceso. Cómo se puede apreciar, la calle o vía pública es el lugar donde se cometen la mayor parte de los robos de vehículos (52,3%), seguido por el domicilio con un 44,1%. Cabe destacar que en este conjunto de datos se ha privilegiado el análisis de los robos de vehículos particulares no comerciales, ya que han sido descartados valores catalogados en la base de datos como “Otros”, donde se sospecha que podrían encontrarse lugares como: bodegas, carreteras, estacionamiento privado, etc.

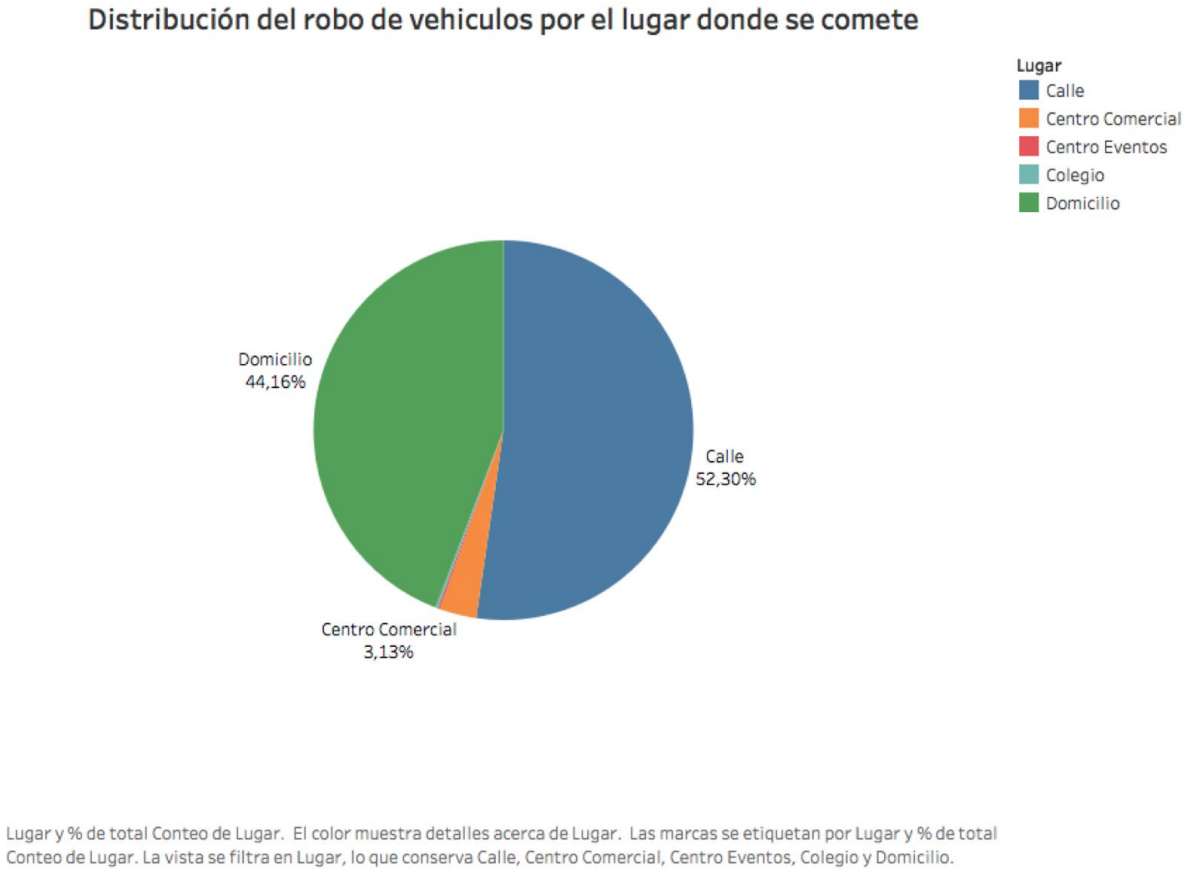

Figura 37. Distribución del robo de vehículos por lugar donde se comete

En la Figura 38 se desgloso por regiones el conjunto presentado en la Figura 37, relevando que existe una clara diferencia entre la Región Metropolitana y el resto de las regiones. Cabe destacar que la Región Metropolitana registra el 83% de los robos en la base de datos utilizada.

En la Región Metropolitana, el 79% de los sucesos se registra en la calle o vía pública, lo que da luces de cuál es el lugar al cual deberían focalizar los esfuerzos por frenar estos ilícitos. En cuanto a las otras 14 regiones, se produce un efecto de inversión respecto a la Región Metropolitana, registrándose un 69% de los ilícitos en el domicilio y sólo un 28% en la calle o vía pública.

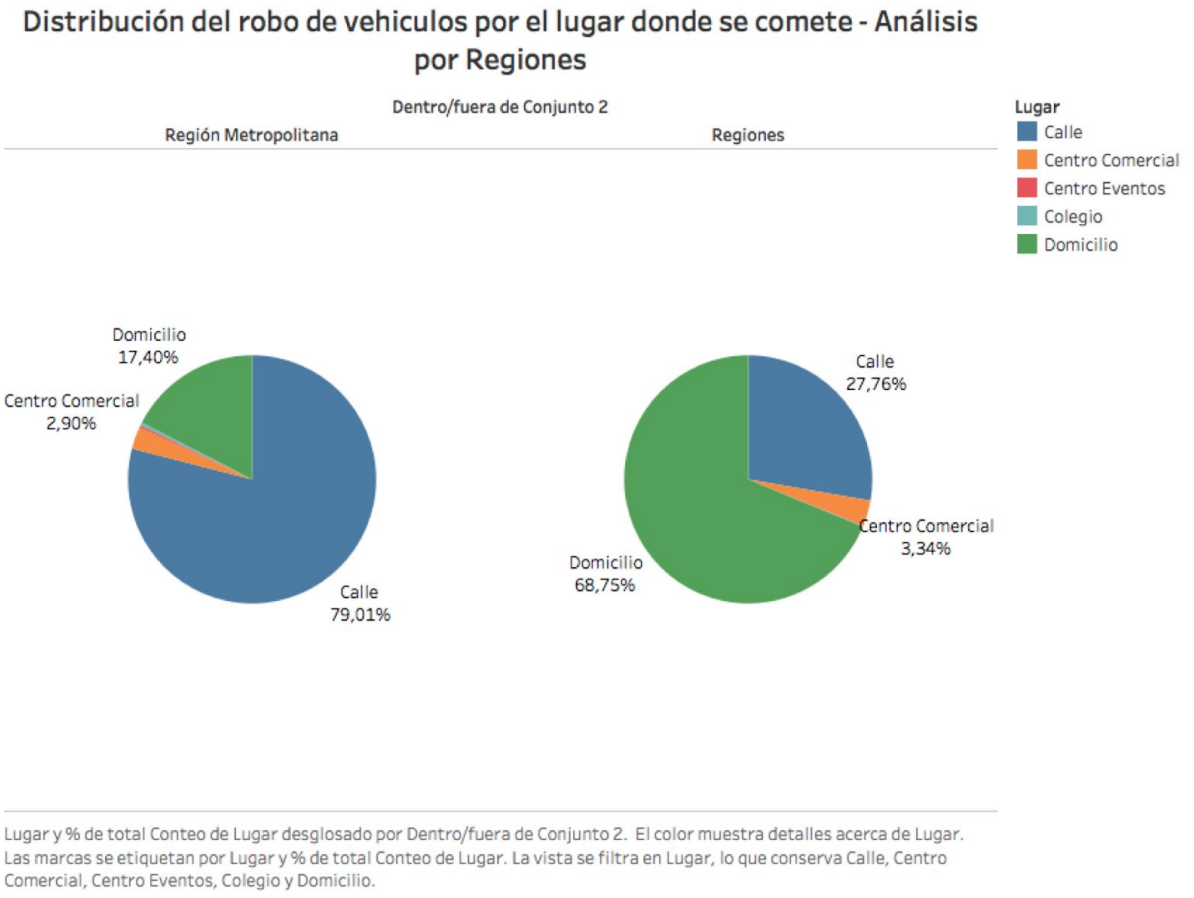

Figura 38 Análisis por regiones de la distribución del robo de vehículos por lugar donde se comete

En la Figura 39 se muestra el análisis que se realizó en la Figura 37 pero considerando la distribución temporal de los robos de vehículos. Cómo se puede apreciar, en Chile el robo de vehículos en la calle o vía pública ha ido en aumento, mientras el robo en domicilio ha ido disminuyendo. Dado que estas dos componentes registran más del 96% de los ilícitos analizados, se puede concluir que estas alzas o disminuciones son efectos significativos en el robo de vehículos.

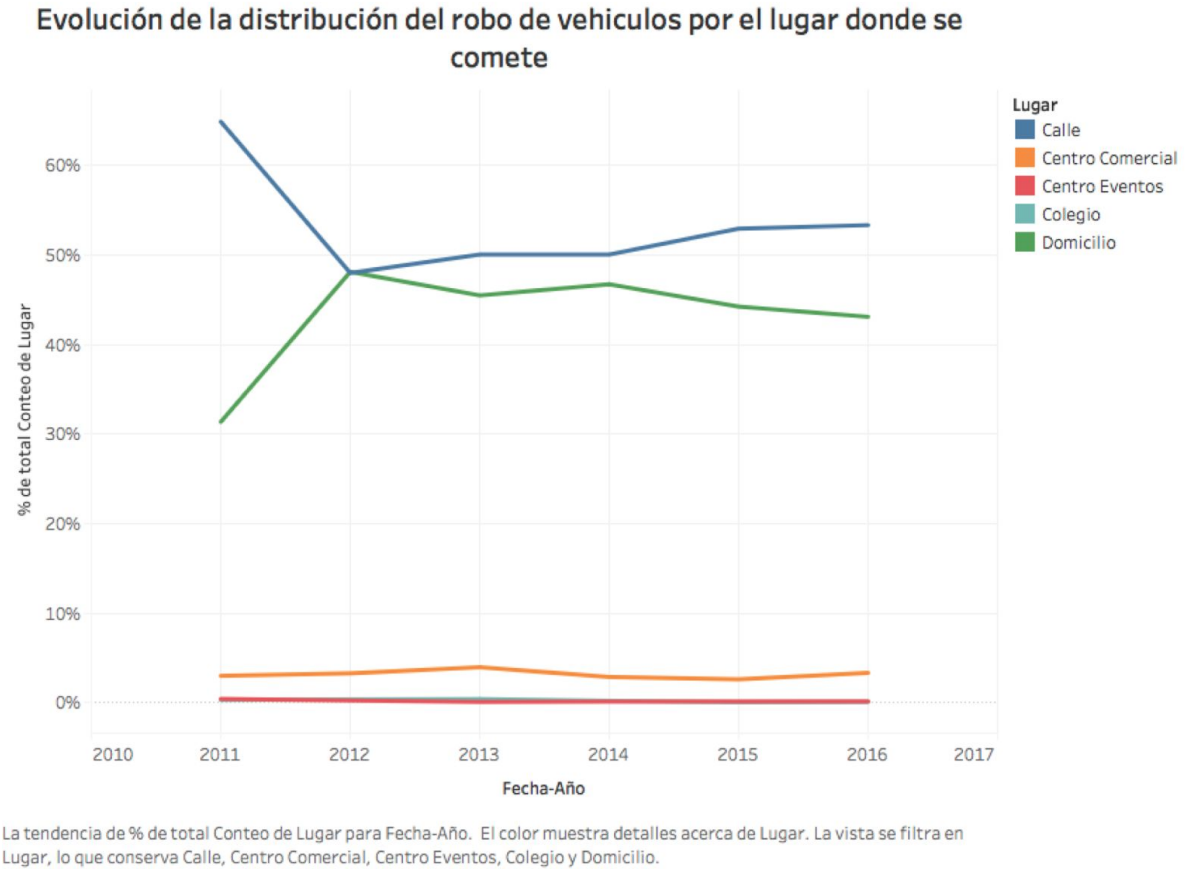

Figura 39 Evolución de la distribución del robo de vehículos por lugar donde se comete

En la Figura 40 se muestra el mismo análisis que se realizó la Figura 39, pero considerando la segmentación entre la Región Metropolitana y el resto de las regiones. Esto es consistente con los resultados anteriores, donde se ha señalado que las proporciones de robo en la calle vs el robo en el domicilio poseen una inversión si se considera el análisis por regiones. Más aún, en la Figura 40 se vislumbra que el robo en la calle o vía pública va en aumento en la Región Metropolitana, registrando un máximo del 84% el año 2014 y cerrando con un 81% el año 2016, mientras que el robo en la calle va en disminución en regiones, alcanzando un mínimo del 23% el 2014 y cerrando con un 30% el año 2016. Así mismo, se puede apreciar que el robo en el domicilio ha disminuido en los últimos 6 años en la Región Metropolitana, alcanzando un mínimo del 13% en el año 2014 y cerrando con un 15% en el 2016.

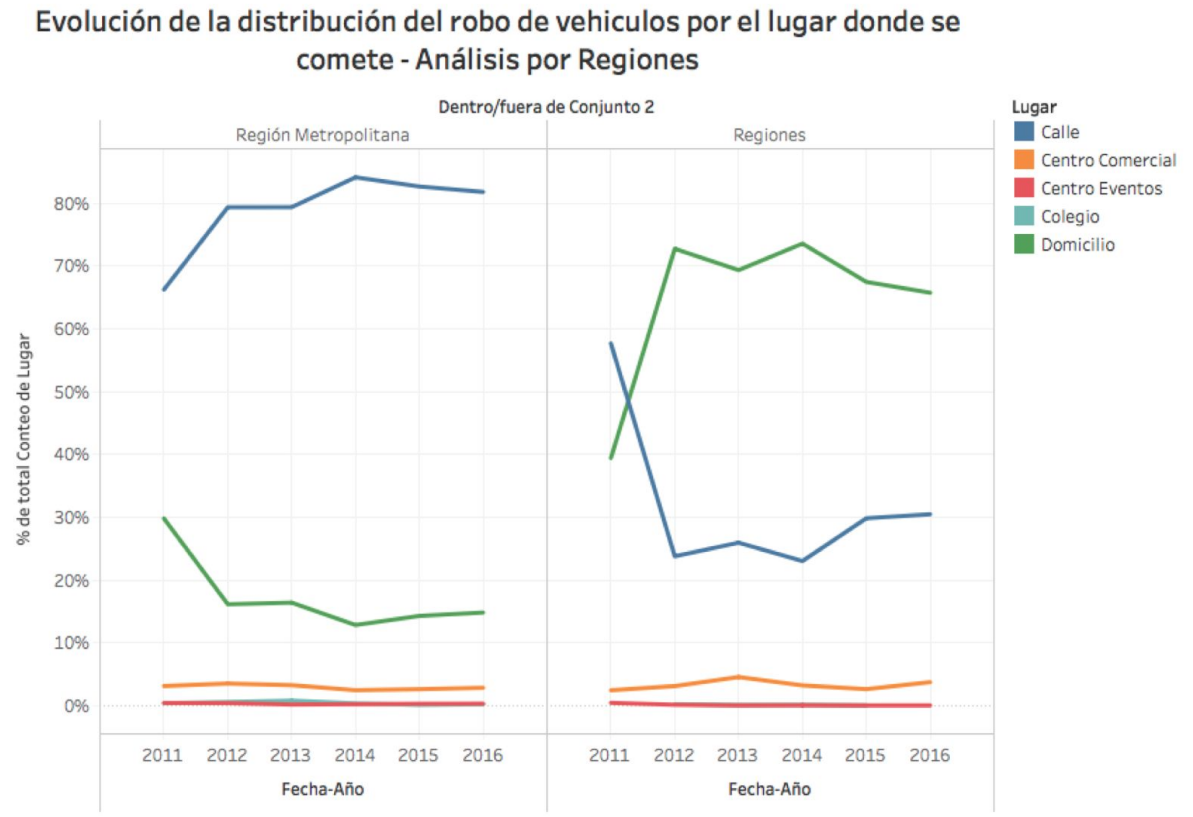

Figura 40. Análisis por regiones de la evolución de la distribución del robo de vehículos por lugar donde se comete

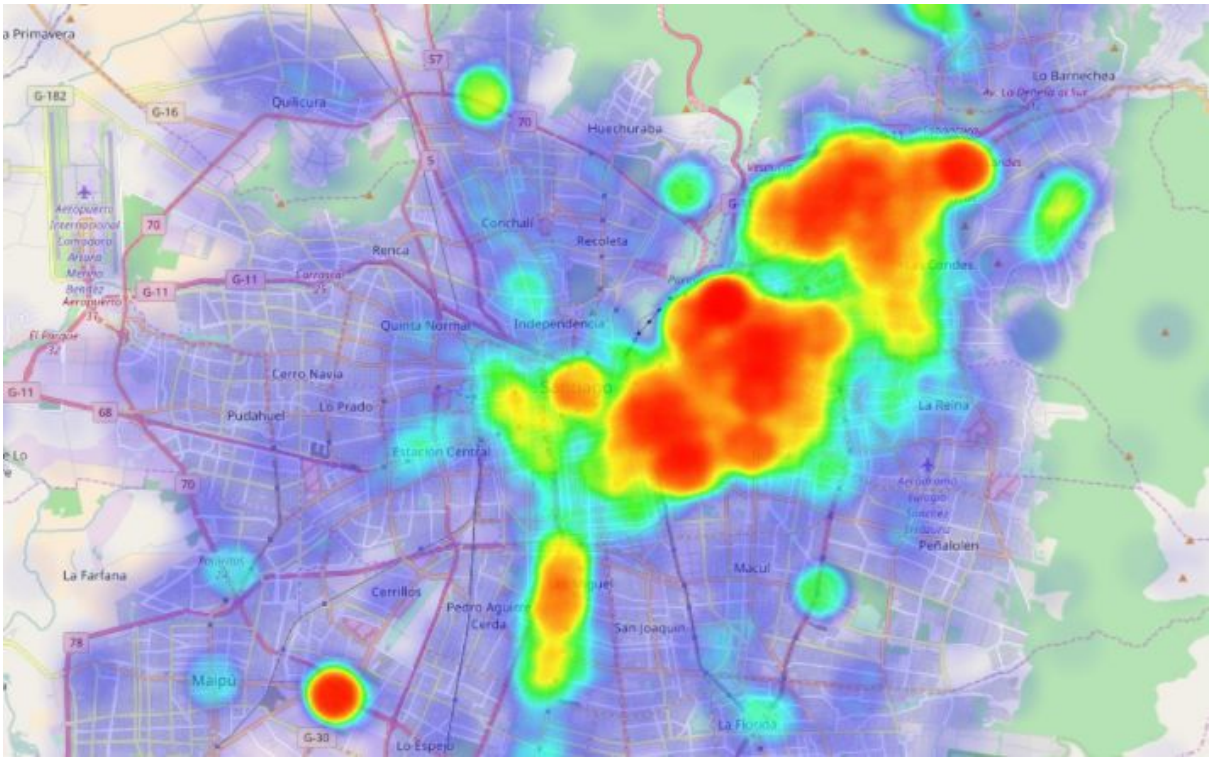

Figura 41. Heatmap del robo de vehículos por comuna de la Región Metropolitana

6. PRINCIPALES RESULTADOS

Twitter ha mostrado correlación con los robos reportados en AACH-DERV, lo cual es bueno, ya que es información que se obtiene en tiempo real, mucho antes de tener reportes de siniestros o denuncias en Carabineros. Adicionalmente, los datos de Twitter proveen información nueva y complementaria a la de AACH-DERV, ya que tienen representatividad sobre un parque automotriz más amplio que el de las compañías aseguradoras. Esta información adicional y novedosa presenta posible valor comercial para las empresas aseguradoras, ya que permite realizar estimaciones de siniestros sobre marcas de autos que ellos no tienen en sus registros (como por ejemplo autos Chinos) y taxis. Esto puede sugerir nuevos mercados y estrategias de negocios.

Al saber que existe una relación fuerte entre los datos de Twitter y los datos reales de siniestros, se puede explotar la información enriquecida de los datos generados por usuarios. Parte importante del valor encontrado en los datos de las redes sociales son sus características espacio-temporales. Es decir, poseen información valiosa en términos de la temporalidad de los robos (horas/días de la semana), que en los reportes no suelen ser muy fidedignas, ya que son imputadas a rangos fijos. También los datos de las redes sociales, poseen mucha información sobre tendencias de desplazamiento de las personas, a pesar de que los datos de Twitter en Chile tienen muy poca información de GPS (alrededor del 2% de los mensajes tienen esta información. Lo cual no hace adecuado esto para trabajar en la detección de patrones espaciales. Esto se ha mejorado utilizando minería de texto para extraer y geolocalizar lugares mencionados en lenguaje natural en el texto de los mensajes y/o en los perfiles de usuarios. Esto permite estudiar ubicaciones y puntos de interés, como lugares y comunas.

El valor de la información de Twitter está directamente relacionado con la cantidad de datos, ya que se extraen estadísticas y datos agregados. Los datos individuales en esta fuente son muy ruidosos y por eso se deben trabajar como un conjunto.

El ruido observado tiene que ver con la ambigüedad en el uso de algunas palabras, que pueden utilizarse dentro del contexto de un robo de autos como también en informaciones que no tienen relación alguna. Esto puede ser mitigado utilizando técnicas de minería de datos que son robustas al ruido, pero requieren volúmenes del orden de miles de mensajes.

El análisis de los **medios digitales de noticias** escogidos muestra que esta fuente de datos no es significativa para la detección de patrones de robos de automóviles. Tampoco provee información valiosa para el apoyo de esta tarea. En primer lugar, del total de portales noticiosos disponibles, sólo unos pocos permiten la extracción de la información digital; los documentos asociados son eliminados luego de un tiempo, lo que no permite realizar un análisis retrospectivo. Por otro lado, menos del 1% de noticias cuenta con información relevante para el proyecto, tal como marca del vehículo, fecha y lugar del robo.

Las razones que guían la publicación de noticias sobre robos de automóviles por parte de los medios digitales, no conducen a la disponibilidad de una fuente de datos apropiada para

el proyecto. En efecto, al comparar la información de los documentos extraídos que hacen referencia a este tipo de robo con la base de datos de la AACH-DERV, no existe correlación positiva significativa ni tendencias similares. Por ejemplo, no necesariamente la cantidad de robos, o sus fechas reales, se ven reflejadas en el número o fecha de aparición de las publicaciones. Esto descarta incluso que las noticias puedan ser utilizadas como una especie de termómetro del verdadero estado de este delito.

Lo anterior, sumado a que no hay información específica del vehículo y la poca disponibilidad de datos, determina que a diferencia de los datos Twitter no se verifique con esta fuente una línea de tiempo que permita alguna especie de trazabilidad desde que se produce el robo del automóvil hasta su hallazgo. Por lo tanto, se propone descartar a los medios digitales de noticias como fuente de datos para el observatorio.

7. REFERENCIAS

[1] Arizona Criminal Justice Commission. (2004). Arizona Auto Theft Study.

[2] Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., & Dufour, B. (2016). Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. International Journal of Agricultural and Environmental Information Systems (IJAEIS), 7(3), 1-20.

[3] Block, C. R. (1995). STAC hot-spot areas: A statistical tool for law enforcement decisions. In Crime analysis through computer mapping. Washington, DC: Police Executive Research Forum (pp. 15-32).

[4] C. B. Block, M. Dabdoub, and S. Fregly. (1995). Crime Analysis Through Computer Mapping. Police Executive Research Forum, Washington D.C.

[5] C. McCaghy, P. Giordano y T. Henson. (1997). Auto Theft: Offender and Offense Characteristics. Criminology, vol. 15, pp. 367-85.

[6] C. R. Block. (1996). The geo-archive: An information foundation for community policing. Unpublished report.

[7] Cameron, J. (2001). Spatial analysis of crime in Appalachia. United States Department of Justice.

[8] Chang, W., Ku, Y., Wu, S., & Chiu, C. (2012). CybercrimeIR–A technological perspective to fight cybercrime. Intelligence and Security Informatics, 36-44.

[9] Chen, H., Atabakhsh, H., Tseng, C., Marshall, B., Kaza, S., Eggers, S., Violette, C. (2005). Visualization in law enforcement. In CHI'05 extended abstracts on Human factors in computing systems (pp. 1268-1271). ACM.

[10] Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., & Schroeder, J. (2003). COPLINK: managing law enforcement data and knowledge. Communications of the ACM, 46(1), 28-34.

[11] Chen, P. S. (2008). Discovering Investigation Clues through Mining Criminal Databases. In Intelligence and Security Informatics (pp. 173-198). Springer Berlin Heidelberg. ISO 690.

[12] D. Gildea and D. Jurafsky. (2002). Automatic labeling of semantic roles. Computational linguistics, vol. 28, no. 3, pp. 245–288.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. (2003). Latent dirichlet allocation. Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022.

[14] D. M. Blei. (2012). “Probabilistic topic models,” Communications of the ACM, vol. 55, no. 4, pp. 77–84.

[15] Dazinger, S. (1976). Explaining urban crime rates. Criminology, 14(2), 291-296.

[16] G. Suresh y R. Tewksbury. (2013). Locations of Motor Vehicle Theft and Recovery,» American Journal of Criminal Justice , nº 38, pp. 200-215.

[17] H. Copes y M. Cherbonneau. (2006).The Key to Auto Theft, Emerging Methods of Auto Theft from Offenders' Perspective. The British Journal of Criminology, nº 46, pp. 917-934.

[18] Hagenauer, J., Helbich, M., & Leitner, M. (2011). Visualization of crime trajectories with self-organizing maps: a case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots. In Proceedings of the 25th conference of the International Cartographic Association, Paris

[19] Kursun, O., Reynolds, K., Eaglin, R., Chen, B., & Georgiopoulos, M. (2005, March). Development of an artificial intelligence system for detection and visualization of auto theft recovery patterns. In Computational Intelligence for Homeland Security and Personal Safety, 2005. CIHSPS 2005. Proceedings of the 2005 IEEE International Conference on (pp. 25-29). IEEE.

[20] L. E. Cohen y M. Felson. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. American Sociological Review, vol. 44, nº 4, pp. 588-608.

[21] M. M. Bradley and P. J. Lang. (2009). Affective norms for english words (anew): Instruction manual and affective ratings. tech. rep., Technical report C-1, the center for research in psychophysiology, University of Florida.

[22] M. S. Gerber. (2014).Predicting crime using twitter and kernel density estimation. Decision Support Systems, vol. 61, pp. 115– 125.

[23] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. ISO 690.

[24] Morenoff, J. D., Sampson, R. J., & Raudenbush, S. W. (2001). Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. Criminology, 39(3), 517-558.

[25] P. Barclay, J. Buckley, P. J. Brantingham, P. L. Brantingham y T. Whinn-Yates. (1996). Preventing auto theft in suburban Vancouver commuter lots: Effects of a bike patrol. NY: Criminal Justice Press, vol. 6, pp. 133-62.

[26] P. S. Chen, K. Chang, T.-P. Hsing y S. Chou. (2006). Mining Criminal Database to Finding Investigation Clues - By Example of Stolen Automobiles Database. Lecture Notes in Computer Science , vol. 2971, pp. 91-102.

[27] Phelps, H. (1928). Frequency of Crime and Punishment. Journal of the American Institute of Criminal Law and Criminology, 19(2), 165-180. doi:10.2307/1134637

[28] Phelps, H. (1929). Cycles of Crime. Journal of the American Institute of Criminal Law and Criminology, 20(1), 107-121. doi:10.2307/1134729

[29] Piskorski, J., Tanev, H., Atkinson, M., Van Der Goot, E., & Zavarella, V. (2011). Online news event extraction for global crisis surveillance. In Transactions on computational collective intelligence V (pp. 182-212). Springer Berlin Heidelberg.

[30] R. A. Bolla. (2014). Crime pattern detection using online social media. Missouri University of Science and Technology.

[31] R. Clarke y P. Harris. (1992). Auto Theft and its Prevention. Crime and Justice: A Review of Research , vol. 16.

[32] R. Clarke. (1991). Preventing Vehicle Theft: A Policy Oriented Review of the Literature,» Scottish Home and Health Department.

[33] R. Siddarth Shankar. (2011). Visualization of the sentiment of the tweets," Master's Thesis, North Carolina State University, Raleigh, NC.

[34] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 1631–1642.

[35] R. V. Clarke. (1989). Theoretical Background to Crime Prevention through Environmental Design (CPTED) and Situational Prevention. pp. 13-20.

[36] Ratcliffe, J. H. (2004). Crime mapping and the training needs of law enforcement. In Crime and Technology (pp. 111-127). Springer Netherlands.

[37] S. Herzog. (2002). Empirical Analysis of Motor Vehicle Theft in Israel, 1990-97.The British Journal of Criminology , vol. 42, pp. 709-728.

[38] S. Sharma, P. Maddili, P. Bankar, R. Kamble y L. Deshpande. (2014). Implementation of Data Mining for Vehicle Theft Detection using Android Application,» International Journal for Research in Emerging Science and Technology, vol. 1, nº 6, pp. 33-37.

[39] Sharef, N. M., & Martin, T. (2015). Evolving fuzzy grammar for crime texts categorization. Applied Soft Computing, 28, 175-187.

[40] Suresh K. Lodha and Arvind K. Verma. (2000) . Spatio-temporal visualization of urban crimes on a GIS grid. In Proceedings of the 8th ACM international symposium on Advances in geographic information systems (GIS '00). ACM, New York, NY, USA, 174-179.

[41] Tanev, H., Piskorski, J., & Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. Natural Language and Information Systems, 207-218.

[42] The European Institute for Crime Prevention and Control (HEUNI). (1997).Motor Vehicle Theft in Europe. The European Institute for Crime Prevention and Control, affiliated with the United Nations (HEUNI).

[43] Torii, M., Yin, L., Nguyen, T., Mazumdar, C. T., Liu, H., Hartley, D. M., & Nelson, N. P. (2011). An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. International journal of medical informatics, 80(1), 56-66.

[44] US Federal Bureau of Investigation. (1991).Crime in the United States - 1998,»

[45] Punyakanok, V., Roth, D. and Yih, W.-T. (2008). The importance of syntactic parsing and inference in semantic role labeling," Computational Linguistics, vol. 34, no. 2, pp. 257–287.

[46] W. B. Groves y R. Sampson. (1989). Community Structure and Crime: Testing Social-Disorganization Theory, Chicago: The University of Chicago.

[47] Watts, R. E. (1931). The influence of population density on crime. Journal of the American Statistical Association, 26(173), 11-20.

[48] X. Chen, Y. Cho, and S. Y. Jang. (2015). Crime prediction using twitter sentiment and weather," in Systems and Information Engineering Design Symposium (SIEDS), 2015, pp. 63–68, IEEE.

[49] X. Wang and D. E. Brown. (2011). The spatio-temporal generalized additive model for criminal incidents," in Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on, pp. 42–47, IEEE.

[50] X. Wang, D. E. Brown, and M. S. Gerber. (2012). Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on, pp. 36–41, IEEE.

[51] X. Wang, M. S. Gerber, and D. E. Brown. (2012) Automatic crime prediction using events extracted from twitter posts. Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, pp. 231–238, Springer, 2012.

[52] Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. Communications of the ACM, 48(6), 100-107.

[53] Zhang, Y., Dang, Y., Chen, H., Thurmond, M., & Larson, C. (2009). Automatic online news monitoring and classification for syndromic surveillance. Decision Support Systems, 47(4), 508-517.

8.- OTROS RESULTADOS COMPROMETIDOS

8.1. Descripción sintética del avance del o los otros resultados comprometidos

Describa en forma sintética los principales resultados comprometidos y su grado de avance a la fecha (Desde el inicio del proyecto).

Resultados Comprometidos	Fecha de logro comprometida en plataforma	Fecha de logro estimada	Porcentaje de avance a la fecha
Resultados de Protección “Derecho de autor.”	30-12-2018	30-12-2018	0%
Resultados de Transferencia y Negocios “…….”	30-12-2018	30-12-2018	0%
Resultado de Producción Científica “Publicación de 4 artículos indexadas en WOS y 4 artículos presentados en conferencias.”	30-10-2018	30-10-2018	10%
Resultado de Formación de Capacidades “Formación de 4 tesistas de pre-/postgrado; participación de un postdoctorado en la creación del Observatorio de Robo de Vehículos”	30-11-2018	30-11-2018	10%

8.2. Descripción detallada del avance del o los otros resultados comprometidos

En el periodo comprendido en el presente proyecto se avanzó con el trabajo con 3 tesistas de postgrado y un postdoctorado para la construcción del Observatorio. Se partió con la elaboración de publicaciones científicas para cumplir con la producción científica prometida.

Trabajos para la protección de la propiedad intelectual y transferencia partirán durante el segundo semestre del año 2017.

9. OBSERVACIONES Y COMENTARIOS

Hasta el momento avanzamos con el proyecto según la carta gantt propuesta sin mayores problemas o atrasos. Consolidamos los datos de los siniestros de AACH-DERV en una base de datos que se desarrollará hacia el Observatorio final. Usamos datos de Twitter en relación al fenómeno de robo de vehículos con muy buenos resultados preliminares.

Sin embargo las noticias digitales no han dado el resultado inicialmente previsto y se propone descartar este análisis en el futuro desarrollo del proyecto.

Asimismo, se propone concentrar los esfuerzos en el robo de vehículos sin considerar explícitamente el robo de accesorios de vehículos. Uno de los argumentos es la escasez de los datos respectivos en redes sociales.

Durante el resto del año 2017 avanzaremos en relación a la protección de la propiedad intelectual y transferencia.

Respecto de los resultados relacionados a la formación de capital humano y producción científica avanzamos según la planificación del proyecto.